Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 08

Lecture – 35

Tests for Mean Vector with Multivariate Data in Two Samples

Hello friend, welcome to the course Multivariate Procedure with R. You can recall that in the last lecture we had considered the Hotelling T square statistics and we have constructed that just of hypothesis for the mean vector in a one sample data set. In case if you try to recall earlier we had considered the T test and we had developed it for one sample test and two sample test. In one sample test we had only one sample and we conducted the test of hypothesis like H0: mu equal to mu0. So, that was done in the last lecture and then again in the univariate case we also developed the two sample T test in which we were trying to compare the population means of two normal population when sigma square is a known, but that is equal from the two populations. Now the same thing I am going to do in a multivariate setup.

So, I already have given you this idea that how we have developed the Hotelling T square statistics and how do we find out the one sample test. So, now the same Hotelling T square statistics will be developed in a two sample case in this lecture and I will try to show you using the R software how you can conduct the test of hypothesis for the equality of two mean vectors when the covariance matrices of the two normal population or multivariate normal population are unknown, but equal. So, let us begin our lecture now. So, now, in this lecture we are going to consider for the test of hypothesis for the mean vector from multivariate data in two samples right.

So, we have the same setting what we have considered earlier that we have conducted a random experiment with P random variable and the random vector is defined as a P cross 1 vector like this of a random variable X1, X2, … Xp. And suppose this x1, x2, … xn is a random sample from multivariate normal distribution with mean vector mu and covariance matrix here sigma and the probability density function of this multivariate

normal distribution is given by this expression right. So, now, you are aware of all these things. Now, we try to consider the test of hypothesis for the mean vector when sigma is unknown. So, in case if you recall we are going to develop it exactly on the same lines as we have done in the case of Junie variate random variable right.

So, first we try to observe here random sample of size n1. So, if you try to see how I am writing it because I have here two sample from two different populations. So, I am writing here the two samples here like this xn1 and suppose my second sample is again x1, x2 here up to here x of size n2. So, now, in order to discriminate between the two samples I am writing here in the super script 1 in the first data set and 2 in the second data set. So, that you can identify that we have got a sample either from the population number 1 or population number 2.

$$\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, \ldots, \underline{x}_{n_1}^{(1)}, \quad \underline{x}_1^{(2)}, \underline{x}_2^{(2)}, \ldots, \underline{x}_{n_2}^{(2)}$$

So, that is how I have indicated here that x1 and in the super script I am writing 1 inside the parenthesis like here this right this. So, this be a random sample of size n1 from a multivariate normal distribution with parameters whose mean vector is mu, but it is now indicated here as a mu super script 1 inside the parenthesis and a known covariance matrix sigma which is a positive definite matrix. Similarly, we observe now here a second sample. If you try to understand it is written exactly in the same way as I have written the first sample. So, this x1, x2, … xn2 this is the sample of size n2 and in the super script I am writing here 2 2 like this.

So, that is indicating that the random sample of size n2 is coming from a multivariate normal distribution with parameters mean vector mu2 and covariance matrix sigma greater than 0. And now we are interested in conducting that test of hypothesis for the equality of 2 mean vector that is my H0 is H0: mu1 is equal to mu2 right. So, and you can see here both mu1 and mu2 they are the p cross 1 vectors right. So, now, what we try to do first we try to estimate this mu1, mu2 and sigma from the given sample of data. So, first based on the sample 1 which is here x1, x2, … xn1 with the super script 1 we try to estimate the mu vector mu1 as the sample mean vector like this one.

So, if you try to see this is simply like that a sample mean vector sample of sample mean of each every each and every random variable right. And this is based on the first temperature like this. And sigma is going to be estimated by here this x11 upon n1 - 1 summation alpha goes from 1 to n1 x1 - x1 bar transpose into x1 - x1 right. You have to

be little bit careful the way I am trying to announce it or saying x1 in the same way. For example, when I say x1 here like this so this is x alpha 1 and this is here x1 bar right.

Based on the sample $\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, \ldots, \underline{x}_{n_1}^{(1)}$

- $\underline{\mu}^{(1)}$ **is estimated as** $\underline{\bar{x}}_1 = \frac{1}{n}\sum_{\alpha=1}^{n_1} \underline{x}_\alpha^{(1)}$
- Σ **is estimated as** $S_1 = \frac{1}{n_1-1}\sum_{\alpha=1}^{n_1}(\underline{x}_\alpha^{(1)} - \underline{\bar{x}}_1)'\,(\underline{x}_\alpha^{(1)} - \underline{\bar{x}}_1)$

So, but that we already have discussed earlier that that how to understand it right. As you try to see here both this mu and the MLE of mu and sigma are estimated by x1 bar and so on right. Do you remember that we had discussed the maximum latinin estimation in the case of multivariate normal distribution area. So, the same thing is being applied here. Now, we consider here the second sample and based on that we try to estimate the mean vector mu2 and covariance matrix sigma here like this.

Based on the sample $\underline{x}_1^{(2)}, \underline{x}_2^{(2)}, \ldots, \underline{x}_{n_2}^{(2)}$

- $\underline{\mu}^{(2)}$ **is estimated as** $\underline{\bar{x}}_2 = \frac{1}{n}\sum_{\alpha=1}^{n_2} \underline{x}_\alpha^{(2)}$
- Σ **is estimated as** $S_2 = \frac{1}{n_2-1}\sum_{\alpha=1}^{n_2}(\underline{x}_\alpha^{(2)} - \underline{\bar{x}}_2)'\,(\underline{x}_\alpha^{(2)} - \underline{\bar{x}}_2)$

So, this is again the sample mean vector of the x1 bar x2 bar up to here x p bar and this is based on the sample size sample of size n2 from the second sample. And similarly this sigma is estimated by here s 2 which is 1 upon n2 - 1 summation alpha goes from 1 to n2 x alpha 2 - x bar 2 transpose into x bar x alpha 2 - x bar 2. Now, based on this the two estimates of this sample covariance matrix is here s1 and here s2. We try to find out here the pooled covariance matrix and it is indicated by here capital S which is (n1 – 1) S1 + (n2 – 1) S2 divided by n1 + n2 - 2. And if you try to recall that is the same thing we have done in the univariate case also where we have observed the two samples from two univariate normal populations and we had estimated the sigma square from the two population and then we had obtained the pooled variance.

$$S = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$$

So, the same concept has been extended here and we have obtained the pooled covariance matrix right. Now, based on that we try to give here the Hotelling T square statistics, but now the Hotelling T square statistics has to be defined for the H0 mu1 equal to mu2. Well I am not going to give you here the mathematical details and derivation of this statistics, but the final outcome here is that the Hotelling T square statistics is

obtained here like this which is n1 n2 upon n1 + n2 and then x bar 1 - x bar 2 transpose s inverse x bar 1 - x bar 2. And this statistics follows the f distribution in this format that is n1 + n2 - k + 1 divided by k into T square upon n1 + n2 this follows a f distribution with k and n1 + n2 - k + 1 degrees of freedom under H0 that is when H0 is true. And this is called as this T square distribution with n1 + n2 degrees of freedom.

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

$$\frac{n_1 + n_2 - k + 1}{k} \frac{T^2}{n_1 + n_2} \sim F_{(k, n_1 + n_2 - k + 1)}$$

Well you need not to worry for these things because when you are trying to implement it in the R software then you have to essentially look at the look into the values of p right. And in case if you want to conduct or find the confidence region for this hypothesis statistics like a mu1 equal to mu2 which I can write down here as say mu1 - mu 2 like this. If you want to find out the confidence interval for the parameter mu1 - mu2 then this is obtained here like that that the confidence region for mu1 equal to mu 2 with confidence level 1 - alpha is the set of vectors m such that x1 bar - x2 bar - m transpose as inverse x1 bar - x bar 2 - m should be less than equal to n1 into upon n1 + n2 into T square with k and n1 + n2 - 2 degrees of freedom or this can be written in terms of f distribution also that you try to replace T square with here of statistics or the related of statistics right. So this is how you can find out the confidence region and as I explained earlier because now we have here p random variables. So this is not called that confidence interval but confidence region.

And you know that this confidence interval concept also helps in that it is just of hypothesis. So that is why this is needed here and you will see that in case if H0 is objective then the R software will also give you the outcome about the confidence interval for the individual components right. So and if you recall I have reproduced here the slide where I had explained you that how are you going to use the confidence interval and in terms of hypothesis. So we had learned that if H0 is rejected at alpha level of significance then there exists a 100 (1- alpha) percent confidence interval which is the same conclusion as the test. So suppose my hypothesis H0 theta equal to theta0.

So if the confidence interval does not contain the value theta0 then H0 is rejected right. This is the same concept what we have used in the case of test of hypothesis. So now we come to our aspect that how to conduct the 2 sample it is it is not V1 equal to V2 in the R software right. So this test requires the package MVTests, capital M V T are in the upper

case alphabets and others are in the lower case alphabets. And this is the same package which was used earlier in the one sample test based on the Hotelling T square statistic.

So first you need to install this package and then you have to upload the package using the command library and after that you have to use the command here TwosamplesHT2. So you have to be careful that here this T of Two, S of Samples and H T that is Hotelling T they are in upper case alphabets and 2 here is numeric. So you have to be careful how do you write this thing. So this is the statistics or this is the command which compute the Hotelling T square statistics for 2 independent sample and gives the confidence intervals also right. And so how are you going to use it? So you have to write down here the command TwoSamplesHT2 as I explained you then you have to write down the data which is going to be a data frame usually and then you have to define here a group.

So this group vector is the values of numbers 1 and 2 which are going to indicate that which of the observations are coming from the population1 and which of the observation in the sample coming are coming from population2. And alpha is here the level of significance as we have used it in the earlier and then homogeneity is equal to true this is indicating that the sample covariance matrices are homogeneous that we have discussed that if the 2 covariance matrices are from where the of the multivariate normal distribution from where the sample have been obtained if they are equal then this is called as a homogeneity. So we have to use here homogeneity is equal to true and if you try to use homogeneity is equal to false then that will be a case that the covariance matrices of the populations are unequal. And in that case the first you have to conduct another test and the homogeneity of covariance matrices can be investigated using the box function. Well I am not going to discuss here, but if you wish you can do it right.

So now this function is going to compute here the the protein key square statistics about the equality of two mean vectors and when H0 is rejected then this command will compute the confidence interval for all the variables to determine the variables which are going to affect the decision of rejecting the null hypothesis right. So this is how the outcome will look like the outcome will have 8 elements the value will be here ST2 which is the value of protein T square test statistics then F value which is the value of F statistics degrees of freedom the degrees of freedom related to the F statistics then P value will be there to conduct the test of hypothesis then there will be Ci which is the confidence interval. So, the lower and upper limits of the confidence intervals for all the variables then the value of alpha that is the level of significance or the confidence coefficient will be given and then we will have here descriptive 1 and descriptive 2 components which are going to give you the descriptive statistics for the observation from the first and second group. So, if you try to see this is the similar outcome what we

have used in the case of one sample Hotelling T square statistics right. So, let me try to take a similar example what we have taken in the one sample Hotelling T square test.

So, means earlier we had the data only on one sample now we will be considering the data on the two sample. So, suppose the data on10 persons in two group is obtained on three variables ages in years, weights in kilograms and hemoglobin level in hb right. So, earlier you have considered the same variable and now I am considering here two different population from where we are going to obtain two different samples. So, the data here in the group 1 it is coming from normal mu1 and covariance matrix sigma. So, there are here person number 1, 2, 3, 4, 5 here you can see.

So, for the person number 1 the age is 10, weight is 35 and hemoglobin level is 11. For the person number 2 the age is 15, weight is 40 and hemoglobin level is 12 and similarly for the third, fourth, fifth persons also we have got the appropriate data. Similarly, in the group 2 the samples are drawn from normal mu2 sigma and here also we have observations on 5 person and for example, the person number 1 in the group number 2 has age 12, weight 33 and hemoglobin level here is 10 and similarly the person number 2 has got age 3 years with 42 kg and hemoglobin level at 11 right. So, now we have here two samples and we want to test here the hypothesis H0 mu1 equal to mu2 at a 5 percent level of significance when sigma is unknown right. So, yeah first I have to give all the data.

So, what I am trying to do here if you try to see I have defined here a variable age 1, 2. Age 1, 2 means if you try to see here the if I try to highlight it in say black color the ages of the persons in group number 1 they are here like this and the ages of the person in the group number 2 they are here like this and now I am trying to join them together and all this data is written here as a data vector. So, if you try to see here these 5 values they are coming from group 1 and these 5 values are coming from group 2 right because this is how you have to give the data right and the same thing I have done for the data on weight and hemoglobin level that first 5 values they are from group 1 and the remaining 5 values they are from group 2 and hemoglobin level also we have the similar conclusion that these 5 values are from group 1 and these 5 values are from group 2 right. And now I try to create a matrix of this data set by here M equal to 10 and col equal to 3 and data is given here by age12, weight12 and hb12 and byrow equal to F. Now, I try to define here a group because if you try to look into this data set age12.

So, you have to inform the R software that ok first 5 values are coming from the population1 and they belong to sample 1 and the remaining 5 values they belong to

sample 2. So, for 1 for this I define here a variable here g like that first 5 values are going to be 1, 1, 1, 1, 1 and the remaining 5 values will be of 2, 2, 2, 2, 2 right. So, for example, if I try to explain you in this data set age12. So, for the first observation I will say here value 1, then 1, then 1 and for first 5 values I have the values here 1 for the variable g and for the remaining 5 observations we have the value here 2 right to indicate that they are coming from group 2.

So, this is how I try to define here my here variable g, then I use here the the package library and we test and then I try to give here the command TwoSampleHT2, data m that is coming from here group is equal to g that is coming from here alpha is equal to 0.05 and homogeneity is equal to true right. And now you will see here actually this outcome will come here like this this will be the data, but I, but since I cannot make it in a single slide. So, I have divided it into 2 columns and I will try to explain you how these things are coming. So, if you try to see here this is value here HT2. So, this is the value of Hoteling p square statistics which we have defined earlier.

This is the value of here f statistics. This is here the degrees of freedom of the statistics the the p value is here like this 0.01 right. So, now, you can see here if you try to see here take here alpha is equal to 0.05 and p value is coming out to be 0.01 which is less than alpha. So, your H0 is rejected and when H0 is rejected then this command will compute the confidence interval for the 3 variables right. So, if you try to see here that these are the actually this part is the confidence interval right for the 3 variables which is indicated by here Ci right. So, these are the lower limits, these are the upper limits and then it is deciding whether the variable is important or not right. So, these are the confidence interval for the first variable you have to read it like this first variable, then second row is for second variable and third row is given the confidence interval for the third variable.

And and how to define this for second and third variable? You have to just see how you have defined your here data say age, weight and say hemoglobin level. So, the first variable will be for about age, second will be for weight and third will be for hemoglobin level right. Then you have here the value of level of specific and alpha, then we have the descriptive statistics for sample 1 right. It is something like the means of the 2 variable and the standard deviation of the 2 variables right. This is for the age, this is for the weight and this is for the hemoglobin level.

And the same thing is here is for the second sample for sample 2 right. So, these are the means of ages right, this is for the weight and this is for here HB level. So, these are the sample means of for the observation in the 3 variables. So, essentially if you try to look at these values they are trying to give you the values of the sample mean vector. And then

similarly here the second row is the standard deviation of the variables 1, 2, 3 that is age, weight and hemoglobin level.

So, this is how you can see that you can conduct the Hotling T-squared statistics for the 2 sample case without any problem. And yeah before I try to go to the part of here let me try to give you an idea the concept about Mahalanobis distance right. That is an important terminology and in multivariate statistics and this is defined as like this mu1 - mu2 vector sigma inverse mu1 - mu2. So, this is mu1 - mu2 transpose. So, this is going to be of order 1 by p, sigma is of order p by p and mu1 - mu2 is of order p by 1.

$$\Delta^2 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

So, this delta square is going to be a scalar quantity of order 1 by 1 right. So, this quantity actually measures the distance between the 2 normal population normal mu1 sigma and normal mu2 sigma right. But again this is based on the population value mu1, mu2 and sigma. So, what we try to do that we try to obtain sample norm first population normal mu1 sigma and from second population normal mu2 sigma and then we try to estimate mu mu1, mu2 and sigma and we try to obtain the sample version of this delta H square right. So, this statistical distance or the Mahalanobis distance between the 2 sample points say x which is consisting of the observations on p variable taken as p is p and y is the sample from the second population normal mu2 sigma having the observation from p variables y1, y2, … yp and they are in the p dimensional space Rp that is understood.

$$d_s(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})' S^{-1} (\underline{x} - \underline{y})}$$

This distance this or this is statistical distance or this Mahalanobis distance is defined as the positive square root of of x - y transpose S inverse x - y and S is the limit of sigma right. So, this is how you can estimate it and if you try to understand there is a close connection between the Mahalanobis distance and the test of hypothesis for the 2 sample mean vector. So, we have this considered the null hypothesis about the equality of 2 mean vectors as s H0 mu1 is equal to mu2. Now, if you try to write down here this quantity here as say mu1 - mu2 like this then this H0 becomes something like mu1 - mu2 is equal to 0 right. And if you try to look at the definition of delta square in case if the mu1 and mu2 are equal to 0 then delta square becomes 0.

So, this hypothesis can also be written as H0 delta square is equal to 0. So, sometime you will see that if in test of hypothesis considered delta square equal to 0 and which is the same thing as that test of hypothesis for the equality of the 2 mean vectors. So, mu1 and mu2 when they are same that equivalent to saying that delta square is equal to 0 and

the same that we can take for testing H0 mu1 equal to mu2 can be used for testing the hypothesis H0 delta square equal to 0 right. So, now let me try to show you this example on the R software. So, you can see here I will just try to copy and paste the entire thing.

Now, you understand it is not a very difficult thing for you and I already have uploaded the packages and we test right. So, if you try to see here I have given here the H, weight and hemoglobin level of these template data, then I have created here a data M matrix and this is here G. So, if you want to see here I can show you here that how the data M looks like it will be like this and G will be here like this right. And now if you if you try to use here this command for the two sample test TwoSampleHT2, then it is coming out to be here like this right.

Let me try to give you this idea one by one. So, this is the value of here Hotelling T square statistic, this is the value of statistic, these are the degrees of freedom, this is here the p value, then we have here the confidence intervals, then we have here the value of alpha level of significance, then we have the values of means and standard deviation for the variables in the sample 1 and descriptive 2 similarly has the values of sample means and standard error from the sample number 2 data and then it is about the test what we have used and its class that is MV test and list right. So, now let us come back we come to an end to this lecture and you can see here that in this lecture we have considered the 2 sample test for the 2 mean vectors which are coming from multivariate normal population, their mean vectors are different, but their covariance is matrix are unknown, but equal. So, and if you ah recall ah that the way we have developed the H0 mu1 equal to mu2 in a 2 sample test in a univariate case, this is almost the same like what we have done there. The only thing is this now we are trying to handle the data vectors in a multivariate setup. So, all this calculation formulation etcetera they will be changed and so that has happened here.

So, now it will be actually good if you try to review the earlier lecture on the 2 sample test and try to see what I have done in this in this lecture also and that will make you very confident and comfortable that how the univariate topics are converted into a multivariate topics. Well it is not a straightforward because for example, the way the distribution of Hotelling T square statistics is found that is quite different than the way other distributions are find, but the basic concepts are the same. So, now my request is that you try to take some data set and you try to practice it in the R software. My suggestion will be that you try to generate 2 samples from 2 different multivariate normal population and try to keep their mean vector almost close or say different try to change the covariance matrix in all the cases right and then try to see that how it is happening when it comes to

the decision from that test of hypothesis. And you will see that the variability in the covariance matrix is also affecting in some way to the test of hypothesis.

So, this type of practice will give you an experience or a hands on experience that how the data speaks and how you have to understand. So, you try to practice it and I will see you with the next topic on scaling in the next lecture till then goodbye. Thank you.