Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 08

Lecture – 34

Testing of Hypothesis: Tests for Mean Vector with Multivariate Data in One Sample

Hello friends, welcome to the course Multivariate Procedures with R. Now, you can recall that in the last couple of lectures, we have conducted the test of hypothesis for the mean in a univariate setup with one sample and two samples. Now, from this lecture and in the next lecture, we are going to understand how we can conduct the test of hypothesis about the mean, or equivalently, the mean vector in a multivariate setup. So, what will be the difference between a univariate setup and a multivariate setup? In a univariate setup, we had only one variable, and we considered only one variable at a time. For example, if there are three variables - height, weight, and age in the case of a univariate setup, we considered only one variable at a time, say height. We collected data on the height of the people and then conducted a test of hypothesis H0: μ equal to some height. Similarly, if I wanted to conduct the test of hypothesis on age, then we collected a sample on the age of the people and conducted the hypothesis H0: μ equal to some age.

But now, I want to work in a multivariate setup, which means I want to construct a random vector where height and weight are both included. This will follow a multivariate normal distribution with a certain mean vector, say μ, and covariance matrix Σ. Now, Σ will have one more component besides the variances of the two variables: it will have σ_1^2, σ_2^2, and the covariance between height and age. So, how do we conduct that test of hypothesis in a multivariate setup? That is what we are going to understand. In this lecture, we are going to consider a one-sample setup. That means we want to conduct a test of hypothesis like H_0: μ equal to μ_0 under two conditions: when Σ is known and when Σ is unknown. Then, in the next lecture, I will extend it to two-sample cases.

That means H0: μ_1 equal to μ_2. Now, in the multivariate setup, μ, μ_1, μ_2, x_1, x_2, ..., x_n will be vectors. So, this is what you have to keep in mind, and I believe that you remember what we have done in the case of multivariate normal distribution. So, let us begin our lecture and try to understand the test of hypothesis in one sample from a normal population in a multivariate setup. Now, we are going to consider the test for the mean vector in multivariate data in a one-sample case, where the data comes from a multivariate normal population.

So now it is important for you to remember whatever we have done in the case of a one-sample test. Right. In the one-sample test, you can recall that we had considered two types of setups: when sigma square was known and when sigma square was unknown. So, when sigma square was known, the proper test statistic followed a normal distribution, and when sigma square was unknown, it was estimated from the sample value, then it followed a t-distribution, right? So, now we consider here a random experiment which has got here where the variable is a random vector, meaning the variables are in a multivariate setup.

This random vector is defined as x_ (the underscore sign indicates that x is a vector quantity). Suppose this is a vector of order p x 1, like this: x1 is the first random variable, x2 is the second random variable, and xp is the pth random variable. For example, x1 could be height, x2 could be age, xp could be weight, etc. We can write this as [x1, x2, ..., xp]^T. Now, we observe the values from here, and let x1, x2, ..., xn be a random sample in a vector setup. So, now, I would like to have your attention: when I say x1, x2, ..., xn in the context of multivariate data, you have to understand it as x_1, x_2, ..., x_n. That means x1, x2, ..., xn are random vectors or vector observations. Saying 'underscore' again and again is difficult and might be irritating for you to hear repeatedly. So, I will simply pronounce it as x1, x2, ..., xn. If it is a multivariate setup, please understand it as a random vector or a vector of observations. If I am talking about a scalar, then understand x1, x2, ..., xn as univariate random variables or scalars. So, now, here you can see I have written x1, x2, ..., xn with underscores. These are the observations on the random vector. So, we have n observations on each of x1, x2, ..., xp.

So, the total number of observations here is equal to actually n into p, p variables, and each variable has got an observation. So, there are np observations in total. That is what you have to keep in mind. Right. So, this sample is coming from a multivariate normal population whose parameters are the mean vector mu and covariance matrix here, capital sigma, which is a positive definite matrix, and its pdf is given here like this.

Right. 1 upon 2 pi raised to the power of p by 2, determinant of sigma raised to the power of 1 by 2, exponential of - 1 upon 2 x - mu transpose sigma inverse x - mu. So, now we are going to first consider the case when sigma square is known in a univariate setup. It is just to make you understand, just to recollect what we have done in the past.

$$\boldsymbol{f(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\right]}.$$

So, in the univariate case, we assume that x1, x2, xn, are coming from a univariate normal distribution, normal mu. sigma square, and we assume that sigma square is known, or the other option is that the sample size is more than 30. In this case, the test statistic was given by X bar - mu upon sigma by root n. So, we try to compute the value of the test statistic using this Zc, which follows a normal 0 1 under the null hypothesis H0 mu equal to mu0. That is, when H0 is true, then Zc will follow a normal distribution with mean 0 and variance 1. Now, we try to extend it to a multivariate.

So, now we have here x1, x2, xn, which are in the form of vectors. So, this is my random sample from a multivariate normal distribution with parameters: mean vector here mu and covariance matrix sigma, which is a positive definite matrix. And suppose we want to test the null hypothesis H0: mu equal to mu0. Now, what does this mean? If you try to see, when I try to write down here mu equal to mu0 like this vector quantity, it is equivalent to I am trying to write down here the mean vector mu 1, mu 2, mu p, which is equal to here some known value mu10, mu20, up to here mu p0.

So, this is going to be the same if mu1 is equal to mu10, mu2 is equal to mu20, etcetera, etcetera, mu p is equal to mu p0, right? That we know. But this is what I mean by this thing, right. So, this mu 0 is a p cross 1 vector of some known values. Now, if you try to consider the multivariate analog of this here Zc, then it is given here by this: x bar - mu0 transpose sigma inverse x bar - mu0.

And this statistic follows a chi-square distribution with p degrees of freedom under H0, when H0 is true, that is, mu is equal to mu0. Well, I'm not giving you here the proof; there is a solid mathematical proof for this statistic also. But here, my objective is not to give the proof but to show you the applications. Right now, if you want to compute this statistic in the R software, then as of now, there is no built-in command. Right? But if you try to look at this statistic, what do you want here? You just need x bar, you just need to specify mu0, and you need to specify this sigma inverse. So, you know that earlier we had discussed that if you want to compute the sample mean vector, then the command here was column means, right.

And if you want to find out the inverse of a matrix, then the command is 'solve'. Right inside the parenthesis, you have to write down the bracket. So, using these commands, I can very easily write down the expression or the syntax for computing this statistic: x bar - mu0 transpose sigma inverse x bar - mu0, right. So, if you try to see, what are going to be the steps? So, the steps are here: I have to specify the data in the form of a matrix, I have to specify the mean vector mu 0, which I am indicating here as mu 0, I have to specify the covariance matrix Sigma, which I am indicating here by capital Σ.

Then, I have to specify the level of significance, then I have to find out n, which is the number of rows in the data matrix x, and p is the number of columns in the x matrix. Now, if I want to find out here x bar, this mean vector, then it is here 'colMeans(x)'. And if I want to find out here the sigma inverse, then it is here 'solve(sigma)'. Now, if I try to write down the test statistic, it will be here n multiplied by the transpose of x bar - mu0— so this 't' here indicates the transpose— and this is here the matrix multiplication, then sigma inverse, and then matrix multiplication, and then here x bar - mu0, and mu0 is also given to us. So, now I can compute this statistic Z2, right.

So, Z2 will follow a chi-square distribution under H0. So, now the p-value will be computed for this statistic by the command 'pchisq', where q is equal to Z2, degrees of freedom is equal to p, and lower.tail is equal to FALSE. If you try to recall, when we discussed the chi-square, t, and F distributions at that time, we had talked about these things. And your decision is going to reject H0 in favor of H1 at a 100 alpha percent level of significance if the p-value is less than alpha. The rule is here the same, whether univariate or multivariate, right? So, now let me try to take here a very simple example to explain to you. Well, I am trying to take here only a couple of variables and a smaller number of observations so that I can show you clearly here what is happening.

So, suppose if I say here that I have five persons, and we call them as person number 1, 2, 3, 4, 5, and we have considered three variables here: age, weight, and the hemoglobin level. So, the age, weight, and hemoglobin level of these five persons have been obtained for example from the person number one the age here is 10 weight is 35 and hemoglobin level is 11 for person number one now for person number two the age is 15 weight is 40 and hemoglobin level is 12 for the person number two similarly for person number three age is 20 weight is 45 hemoglobin level is 13 and this is person number three then for person number four the age is 25 weight is 50 and hb is 14 for person number four and similarly here for person number five the age high weight and hemoglobin level are 30 55 and 15. they are yeah in some proper units And suppose we want to test here the null hypothesis H0 mu equal to mu0 at 5% level of significance. And suppose mu0 is given to be here 25, 45, 13 transpose, right? And suppose sigma is known.

So, just for the sake of simplicity, I am taking here diagonal matrix of here 1, 2, 3. You can take any positive definite matrix here, no issues. My objective here is only to explain you and demonstrate how it is going to work right. So, if you try to see this is the these are my steps what I have taken here in the step number one two and three. I am indicating the data what I have said then in the step number here before I am trying to create here the matrix of this age weight and hemoglobin data. Then I am trying to specify here the mean vector and sigma matrix in the step number 5 and 6 and then I try to define here this here number of rows and number of columns to specify the n and p.

And then I try to find out here the x bar that is the mean vector by this command here bar dot x and then I am trying to find out here this here sigma inverse by writing here all sigma and then in the step number here 11, I am trying to write down here the statistics z2 is equal to here like this and then the p-value is computed by here this command here. The same command at step number 13 and then I will get here the p-value so this is step number 12 and step number 14 will give me the value of the chi-square statistics and p-value, right. So, this is the data set in a neat way now if I try to see the data in the matrix format looks here like this and if you try to take here this here z2 z2 here is like this and if you try to execute it it will giving it is giving you a value here 1574.067 and now at the next step if you try to find out here the p value which is computed over here and p value is coming out to be here zero yeah because it is artificial data So, now you can use here the decision rule that H0 is rejected in favor of H1 at when your P value is equal than alpha, right. So, you can see here it is not difficult. Now, this here P value is here almost 0 and alpha is here 0.05.

So, you can see here this p-value is less than alpha. So, this H0 is rejected. So, the conclusion here is to reject H0, right? Okay, and then you can see here this is the screenshot. So, let me try to show you this thing on our console also. So, I will simply try to copy and paste these things here so that, yeah, I save time. And if I try to do it here, you can see here that the Z2 statistic is coming here like this, and the P-value here is coming out to be like this, right? So, now similarly, you can take some more examples and try to do such exercises.

Now, after this, I will try to consider here the next case when this variance or equivalently the covariance matrix is unknown. So, just for your revision, I am taking the univariate case so that you can recall, and then I will try to extend it to a multivariate case also, right? So, in the univariate case, you remember that we have taken a random sample X1, X2, ..., Xn from normal (μ, σ^2). where μ is unknown, σ is also unknown, the population is normal, and the sample size is smaller, less than or equal to 30. And we had used the test statistic Tc, which was ($\bar{X}$ - μ) / (S / $\sqrt{}$n), that was following a T-distribution under H_0, where the variance S^2 is estimated by S^2, right?

$$T_c = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ **.**

So, now I would like to extend it to a multivariate case, right? So, I am assuming that now these X1, X2, ..., Xn, they are in the form of vectors of values. This is a random sample of size n from a multivariate normal distribution with parameters μ (vector) and covariance matrix Σ, and Σ is unknown. So, now based on this sample X1, X2, ..., Xn, the first step is that we try to estimate here the mean vector by the sample mean vector like this. So, with all the observations, the variable-wise arithmetic mean has been obtained, and we try to get the covariance matrix here by this quantity here, the Σ estimated here, let us say S = (1 / (n - 1)) * Σ (from α = 1 to n) (Xα - $\bar{X}$)^T (Xα - $\bar{X}$).

Multivariate analogue of $T_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ **is**

$$T^2 = n(\underline{\bar{x}} - \underline{\mu}_0)' S^{-1} (\underline{\bar{x}} - \underline{\mu}_0)$$

T^2 **is called as Hotelling** T^2 **- statistics which follows**

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

under H_0: $\underline{\mu} = \underline{\mu}_0$**. It is termed as** T^2 **- distribution with** n **degrees of freedom.**

Right now our null hypothesis is s0 mu is equal to mu0 where mu and mu0 are the p cross 1 vectors so mu0 here is a p cross 1 vector of known values now we try to develop the multivariate analog of this tc which was used in the univariate case so if you try to see here if I try to square this tc this will look like here tc square say x bar - mu0 and it will be your here n then x bar - mu0 and then here this is a small s small s will become something like a matrix so i try to write down here s inverse and this is a scalar quantity so i put here the transpose here like this well this is this looks like as if t square or this t square c has been derived in this way but i would like to inform you all that there is a proper statistical method using likelihood ratio test. We derive the statistics right. So, here this way the way I have explained you it is just for the sake of your understanding.

The same statistics is derived using the likelihood ratio test or the union intersection principle in multivariate analysis. But I am not going to discuss that part here, but I am simply going to demonstrate the use of this statistic. So, the multivariate analog of Tc is here like this. T square is equal to small n into X bar - mu0 transpose capital S inverse X bar - mu0.

And this T square is called here as a Hotling T square statistics. This is also cordless generalized t square statistic. Right. And about that distribution, for example, if you remember, they said Tc was following a t distribution with n - 1 degrees of freedom. Similarly, we have to do some mathematical statistics and then we find that this expression here that is N - K + 1 divided by K into T square upon N.

This follows F distribution with K and N - K + 1 degrees of freedom under H0 mu is equal to mu0. And this is termed as t square distribution with n degrees of freedom, right? And in case if you want to construct the confidence region for this mean vector at confidence level n - alpha, then this is going to be the set of vectors mu such that n into x bar - mu transpose s inverse x bar - mu is less than equal to t square with k n and - 1 degrees of freedom if which is actually equivalent to saying that n into x bar - mu transpose s inverse x bar - mu is less than or equal to n - 1 k divided by n - k into f value at k and n - k + 1 degrees of freedom at alpha level of significance.

$$\boldsymbol{n(\underline{\overline{x}} - \underline{\mu})'S^{-1}(\underline{\overline{x}} - \underline{\mu}) \leq T^2_{k,n-1}}$$

or

$$\boldsymbol{n(\underline{\overline{x}} - \underline{\mu})'S^{-1}(\underline{\overline{x}} - \underline{\mu}) \leq \frac{(n-1)k}{n-k} F_{(k,n-k+1)}\,(\alpha)}$$

So if you can recall that when we considered the univariate case then we also considered we constructed the confidence interval and similarly here also we are considering a multivariate case so we have considered a confidence region it is not a univariate case that the confidence interval is lying on the line, but it is a confidence region because each of this mu 1, mu 2, mu p will have confidence interval. And now if you try to combine them together, it will look like a confidence region, right? So, now if you want to conduct the test of hypothesis in such a case in the R software, then it requires the package MVTESTS, MVTests where MVT, they are in capital, and ests they are in lowercase so you need to install it using this command you have to upload it and then and the command here is OneSampleHT2 and this compute the one sample hotlink t square star six and give the confidence interval right So the use is here like this.

This is the command here. OneSampleHT2. You have to be very careful that here O, S, H and T, they are in uppercase alphabet and 2 is in number. And then you have here data, the value of new 0, level of significance alpha. So, all these values you have to give and this will give you the value of one sample Hotelling T squared statistics.

And when H0 is rejected, this function will compute the confidence interval for all the variables because you know how you are going to take the decision for the test of hypothesis using the confidence interval. Right, so when you see the output of this command, it will have seven elements: HT2, which will give you the value of Hotelling's T-squared statistics; F, which is the value of F-statistics; df, the corresponding degrees of freedom of the F-statistics; p-value, which is the p-value; Ci, which is the confidence interval, that is, the lower and upper limits of the confidence interval obtained for all the variables; and alpha. Alpha is the level of significance. So, these seven outputs will be there in the outcome of this command.

Right. So, let me try to take here the same data which I have just used in the case when sigma is known. Right. So, I have here the same data that we have here: 5%, and then we have obtained the data on their age, weight, and hemoglobin levels. And suppose we want to test the same hypothesis H0: mu equal to mu0 at a 5% level of significance, where mu0 is given by like this.

And sigma is now here unknown. So, if you try to see here, I try to give here the data, I try to create here the data matrix, specify here my mu, and then I compute here, I execute here the command: OneSampleHT2, and then the data matrix here, data M, value of here mu0, value of here alpha. Right, and I get here this type of outcome. Actually, I get here this type of outcome. I can first show you the screenshot here, which I have divided into two parts so that I can explain. So, you can see here this HT2 is giving us the value of Hotelling's T-squared statistics. F value here. This is the value of F-statistics.

Degree of freedom, it is something like F with 3 and 2 degrees of freedom. P-value, which is here like this. And you can see here that the P-value is 0.001, which is less than alpha equal to 0.05. So, that is why H0 is rejected. And so, the confidence intervals are computed here, which are given like this.

So, this is the confidence interval for the first variable, x1. This is the confidence interval for your second variable, and this is the confidence interval for the third variable, right? So, x1 was your weight, x2 was your height, and x3 was your hemoglobin level. Then it gives you here the value of your alpha. Under the descriptive, it gives you the number of observations for each variable, which is here 5. The column means are given here, and then the standard deviations of each variable are given here. That is the descriptive statistics. So, you can see here it is not very difficult to execute. If I try to execute it on the R console here, then it is not that difficult; it is very easy. I try to just copy and paste these commands. First, I have to upload this package, right? So, if you try to see, I already have installed this package on my computer. So, I have to use the library and we test. So, if it is here, then if you try to see here, I try to use all these things over here. And I try to see here, this is here like this.

So, you can see here, if you try to go here, this is the value of the t-squared statistic. This is the value of the F-statistic. This is here degrees of freedom. This is here the p-value. This is here the confidence interval and this is here the alpha value this is here the descriptive statistics that is the number of observation means standard deviation and it is describing here what is this test this is OnesampleHT2 and it this is here the class right with where it belongs so now you can see here it is not very difficult to do it right so now with this description I come to an end to this lecture And you can see that it is not very difficult. It is simply the methodology and the approach is exactly the same what we did in the case of univariate random variable. And even the form of the statistics is look like multivariate generalization of the univariate case. But as I said, they are derived using the likelihood ratio test. And their distribution is found using some mathematical and statistical tools. It is not like by looking at the t-statuses, I will square it and then adjust it. But it looks like so. I just introduce it for your understanding only. You please don't misunderstand it.

And now you can see here doing it in the software is not difficult at all. But the main challenge is how to interpret it. Because now you are coming with a variable. which are more than one in number right i have considered three variables but you can have 20 variables 100 variables so now how to take the correct interpretation of them that is the job of data scientists who are working with the real data and those conclusion they have to collaborate with what is happening in the real life also So my request to you will be you please try to take some data and try to conduct this test of hypothesis.

One simple way out is that now you know how to generate data in a multivariate normal distribution. So, take some mu vector, take some covariance matrix, and try to generate the data with some number of observations. Try to save those observations in the form of a matrix or a data frame so that they can be used in this test, and then try to conduct the hypothesis test. Try to take different values of mu, try various values of covariance matrices, and then observe how these values change. That will give you an impression of how multivariate things behave and how variables behave in a multivariate setup.

So, try to practice it, and I will see you in the next lecture. Till then, goodbye.