Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 07

Lecture – 29

Basics of Tests of Hypothesis

Hello friend, welcome to the course multivariate procedure with R. So you can recall that in the last lecture we had talked about the estimation of parameters and we had seen that how to use the method of maximum likelihood for estimating the unknown parameters which can be univariate or which can be multivariate. Now you have seen that when we try to obtain the estimates of the parameter or the value of the parameter theta based on any sample then as the sample will change the value of the parameter will also change. How? So in this example we had obtained in the case of normal distribution that the maximum likelihood estimator of mean is sample mean. So that means you are going to collect a sample and then you are trying to find out the mean of those observation. Now definitely you can once you take another sample the sample values are going to be changed and then the sample will also change.

So when you try to draw different samples the sample values will be changing. Now how to ensure that which value is correct or which value is the correct value of the theta or which value is representing the value of theta correctly. The representative value of theta is the value of theta which is prevailing in the population not in the sample. So one idea is that if there is not much difference between the values of parameter which are obtained from different samples possibly we can say that okay they are okay and they are representing the good value.

For example suppose if I make here a claim that my age is 5 years will you believe on it certainly no and if I say my age is 20 years you will say no it is not like this but if I say my age is suppose 55 years then possibly you cannot say no. But the question is you do not need to know my age but my question is that when I said a 5 years, 15 years, 20 years then you said no but as soon as I said 55 years you could not say no. So what is

happening that as soon as I say my age is suppose 5 years then you try to compare my structure, my face etc. etc. and you try to compare it with someone who is already 5 years old or you try to compare it with the data which you have stored in your mind about the 5 year old child and there is a huge difference between the two values and that is why you say yeah I am not 5 years old but when I say that I am 55 years old then you try to compare my data with someone who is 55 years older maybe in your family, maybe in your friend circle etc. and then you say yeah my body structure, my eyes, my hands, my skin, hairs on my head etc. they resemble more with a 55 year old person and that is why you believe on my statement.

So if you try to see this whole process was believing or not on a statement and this belief or the decision to believe or not is based on certain data. I say a statement, you try to create another statement in your mind and you try to compare it on the basis of some data set and then whatever is the outcome based on that you try to say whether I am correct or you are correct. If I am correct then you cannot be correct, if I am not correct then you are correct.

So what is happening in between that how you are trying to manipulate this data inside your mind so that you can come to a conclusion that whether I am right or you are right. This is the job which is done by test of hypothesis. What is hypothesis? Hypothesis is only a statement. So whenever we are dealing in statistics all our statistical procedures they are based on the random sample. The random samples vary and consequently the values of those parameters, values of those statistics they also vary and we want to test whether the values which you have obtained on the basis of given sample of data are they dependable or not on the basis of certain criteria.

So now this test of hypothesis can be done on a univariate random variable on a multivariate random vector also, means including the multivariate random matrix also. So now from this lecture and in the next couple of lecture we are going to talk about the test of hypothesis and confidence interval estimation. Why confidence interval estimation? That you will see that the results of test of hypothesis and confidence interval estimation they are interrelated. Once you know the result of test of hypothesis then you can also get the result about confidence interval estimation or by looking at the result of the confidence interval estimation you can also take the conclusion about the test of hypothesis. So based on different types of situation we have different types of test of hypothesis but surely as I say always my objective is not here to teach you the test of hypothesis and confidence interval estimation.

My objective is that I want to give you here the sufficient knowledge about these topic so that when we are trying to go further and when we try to use that concept of test of hypothesis and confidence interval estimation in our multivariate procedure then you understand it. So with this objective in this lecture I am going to give you here a brief overview about this procedure that how do we conduct the test of hypothesis. And in the next lecture we will take specific examples to conduct the test of hypothesis particularly about the mean in the normal population and the mean vector in the multivariate normal population. So let us begin our lecture and try to understand about the test of hypothesis. So in this lecture now we are going to talk about the basic of test of hypothesis, some definitions, some terminologies and how do we conduct it.

So as I explain you that we have noticed that when we try to draw different types of sample they will give us the different values of estimates and the value of these estimates may or may not be the same as of the unknown value of the parameters in the population. So this difference in the estimated value may be due to some random variation or there may be some assignable cause. For example, in the example which I just took about my age when I said my age is 5 years you can say that okay there is a assignable cause, there is a reason and you cannot believe on my statement. But when I said 55 years then you accepted it. On the other hand if I say okay my age is 54 years, 6 months then once again you cannot say that it is not my correct age.

But the difference between 55 years and if 54 years, 6 months or 55 years, 6 months it is only a matter of 6 months which hardly makes any difference on anything. So this variation can be caused, it can be said as if this is a random variation. So the value which we want to know is unknown, that is the unknown parameter. So we want to check if the estimated value differs from some other reasonable value or not. So if I try to translate it in very simple words then we would like to test the closeness of the estimated value with some hypothetical value.

What is this hypothetical value? Suppose in the same example of my age as soon as I say or make a statement that my age is 5 years you always try to assume a value inside your mind that if there is someone whose body structure looks like as of mine the age cannot be 5 years or the age should be more than 5 years or the age could be close to 54 years, 55 years, 50 years and so on. So this is hypothetical value which is existing inside your mind. So what are we trying to do? We try to find out the difference between the estimated value and the hypothetical value. And if this difference comes out to be less then we can expect that this is going to be accepted and we would say that there is not much significant difference between the two values. For example, when I say my age is

54 years, 6 months and 55 years then you say there is not much significant difference between the two values.

And if this difference increases then we can expect to reject it and we would say that there is a significant difference between the two values. For example, if I make a statement my age is 5 years and you make a statement that my age is 55 years then the difference is pretty high and we say that okay there is a significant difference between the two values. But the question now here is how to decide scientifically that the difference between the two values is significant or not. So we try to take the help of this tool or statistical technique of test of hypothesis. So what is a hypothesis? A statistical hypothesis is usually a statement about a set of parameters of a population distribution.

It is called a hypothesis because it is not known whether or not it is true. For example, if I say my age is 52 years now whether my statement is true or not that you do not know. So that is why it is a hypothesis. And a primary problem in such a case is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis. Because as soon as I say my age is 55 years you try to do some calculations in your mind and you try to get an answer yes my age cannot be 55 years or my age can be 55 years because it depends on certain body characteristics.

So what is that procedure? So this is a challenge. So basically when we try to develop such a procedure then we have to make a rule. And this rule is going to be dependent on the random sample what we are observing from a given population. So a test of a statistical hypothesis is a rule or procedure for deciding whether to reject the hypothesis or not. Now in case if I try to take the example of a normally distributed population which has got an unknown mean value suppose here mu and some known variance here 1.

Now in this case variance is known so I am not bothered about it but the mean is unknown so we would like to know its value on the basis of the given sample of data. So now if I try to create here a statement that mu is less than 17 right. That this means it is a statistical hypothesis and we could try to test by observing a random sample from this population whether mu is less than 17 or not. And if the random sample is deemed to be consistent with the hypothesis under consideration then we say that the hypothesis has been accepted right. For example if I say my age is 55 years and you say my age is 55 years 2 months there is not much difference between that two values and who can say that okay so my value is accepted.

And in case not accepted then we say that this hypothesis has been rejected. So accepted and rejected these are the two terminologies that we use to decide the outcome of a test on the basis of given sample of data. Now this test of hypothesis can be conducted in different scenarios based on different types of samples. So when we are working only with one sample to conclude about a value the problem is termed as one sample problem. For example if I try to take a sample from a class to determine that what is the age of the student in that class.

So I can have a sample based on that I can compute the sample mean and then I will conduct the test of hypothesis. So there is only one sample and we want to make a conclusion only about the parameter based on one sample. So this is one sample problem. Now on the other hand if I try to compare the average ages of two different classes then what will I do? I will try to take one sample from class number 1 and say another sample from class number 2 and then I will try to compare their ages on the basis of given sample of data. So now in this case particularly we are observing two samples to make a comparison and a conclusion about the value then this type of problem is termed as two sample problem.

And if I try to extend it when we are working with more than two samples to make a comparison and conclusion about a value the problem is termed as multivariate sample problem. For example if I say a school has 10 sections in say class 10. So now there are 10 sections. So now in an exam suppose I want to compare the average marks of the students obtained in every section. So now there will be section 1, section 2, section 10.

So we will try to draw here 10 different samples from each of the section and then we will have here a problem which is multivariate sample problem where we will be comparing the means of 10 samples. So now the question comes here how to construct the decision rule. So let me try to give you here this brief background that when we construct the test then we simply try to partition the sample space of X1, X2,… Xn into two disjoint subsets for example say C and C star. Suppose this is my here this sample space and I try to divide this into say C and C star. And then we try to obtain here a sample X1, X2,… Xn and then we try to compute C star based on that we try to take a decision that x1, x2,… xn is belonging to suppose here C.

$(x_1, x_2, \dots, x_n) \in C$

Then we suppose we decide that if my sample values are belonging to the region C then we will reject the H0. H0 is a hypothesis which we call as null hypothesis that I will try to discuss soon. And on the other hand in case if I say that x1, x2,… xn are such that this

point belong to C star that means this region here. Then we can say that we shall accept the null hypothesis H0. So based on this classification this C here is called as critical region or the region of rejection that means the sample values are going to lie in this region and we are going to reject the hypothesis.

On the other hand this region here C star this is called here as a acceptance region or the region of acceptance that means if the sample values are going to lie in this region we are going to accept the hypothesis. And in order to avoid any confusion ambiguity we say that there is no common region between C and C star right they are a joint. There is no region where I can say that X1, X2, … Xn will belong to C as well as C star right. So now the question comes here what is the statistical hypothesis in terms of the parameters right. So as I said that hypothesis is a statement about the parameter and this statistical hypothesis has two parts which is called here null hypothesis and alternative hypothesis.

$(x_1, x_2, \dots, x_n) \in C^*$

So what does this mean? Suppose I try to take the same example that I try to make a statement that my age is 20 years and now you have to test it on the basis of sample of data whether my statement is correct or not. In order to compare it you have to create one more statement. Now you create a statement here my age is suppose 55 years. So I want to test my statement that my age is 20 years in comparison to your statement that my age is 55 years right. So that is why I am saying that here we have two parts one is here called as null hypothesis which is denoted by H0 and this is a statement which is to be tested or that we want to test and other part is alternative hypothesis which is denoted as here H1 or say Ha and it is just opposite to the null hypothesis and this is a statement against which the null hypothesis is tested right.

Both this null and alternative hypothesis they are mutually exclusive and only one of them can be true at a time. Either H0 can be true or H1 can be true. If H0 is true then H1 cannot be true and if H1 is true then H0 cannot be true right and whenever we are trying to test any hypothesis then we have to take a decision. Now this decision can have different situation and different out.

So let us try to understand it. Suppose in the actual situation there are two option that the null hypothesis can be true or the null hypothesis can be false. Now when you try to talk about the decisions what are the possible decisions. One possible decision is that we try to accept the H0 or we try to do not accept that is reject H0. So these are the only two

options whether you are trying to accept or not. So now if you try to see what are the possible decisions.

So let me call it say here box number 1, box number here 2 in this way box number here 3 and box number here 4 right. And let us try to analyze the decision each of the box. So let us try to concentrate on box number 1. So when I am saying that H0 is true and we are going to accept it absolutely no issue there is no error that is what we want to do.

Now you come to box number 2. We are trying to say that H0 is false and we are going to reject it like this here like this right. In that case absolutely there is no issue. So now I come to here now we come to box number here 3. You can see here H0 is here false and we are saying that H0 is accepted that means we are trying to make a mistake. On the other hand if you come to here box number here 4, 4 you are trying to say that H0 is true and you are not accepting it.

So that means we are trying to make here mistake. So you can see here in box number 3 and box number 4 we are trying to make 2 possible mistakes in making a correct decisions. So these are 2 types of errors which are happening in a decision making. So the error which is happening in box number 4 that H0 is true and we are not accepting it this is called as type 1 error. And the error which is happening in box number 3 that H0 is false and we are accepting it this is called as type 2 error.

Decision	Actual Situation	
	H_0 True	H_0 False
Accept H_0	No Error Probability $1 - \alpha$	Type II Error Probability β
Reject H_0	Type I Error Probability α Level of Significance (Critical region)	No Error Probability $1 - \beta$

Now the question here is how to say decide the magnitude about these errors. So we try to compute the probability. So the probability of type 1 error this is called here as say alpha and this is also called as the level of significance and this is the size of the critical region. What is critical region? If you try to see in this slide we have talked about the

critical region here that this is the region where our data lies and we reject the hypothesis or reject the null hypothesis.

Now similarly you come to here box number 3. So box number 3 is also making type 2 error and we want to quantify it. So we try to compute the probability of making type 2 error and this is indicated by here like this. So the type 1 error is indicated by alpha and the type 2 error is indicated by the beta or more precisely the probability of type 1 error is indicated by alpha and the probability of type 2 error is indicated by beta. So these are essentially the terminologies that we try to use in the test of hypothesis. Similarly if you try to look into the box number here 1, the probability of no error is here 1 minus alpha and in box number 2 the probability of making no error is 1 minus beta.

But anyway we will try to concentrate or work with the type 1 error and type 2 error only. But definitely I am not going to give you here more details because when we try to develop our decision rule we have to develop our decision rule in such a way such that the probabilities of these two errors are minimum. And now minimizing both the errors is not possible. So we try to follow the rule which is given by the Neyman Pearson Lemma which is based on the basic assumption that type 1 error is supposed to be more serious than type 2 error. So what we try to do here that whenever we try to develop the decision rule then following the Neyman Pearson Lemma we try to fix alpha and we try to minimize beta that is the type 2 error.

So that is the basic fundamental rule. So I am just giving you an idea and if you try to read from any book or any types of standard test statistics that we are going to use here they have been found using the Neyman Pearson Lemma and they are based on the assumption that type 1 error is more serious than the type 2 error. So now I will just try to give you here the formal definition of different terminologies. So what is the meaning of accepting H0? This means that the difference between sample mean and hypothetical population mean is not significant or whatsoever is the difference this is coming because of the sampling fluctuations only. There is no assignable cause it is only the random variation because of which the different values of estimator from different sample based on different sample they are varying. Now another concept in the test of hypothesis is about p values.

So as I told you that when we try to take a decision whether to accept or not the null hypothesis then we try to decide the critical reason and then using our test we try to decide where it is going whether to the critical reason that is the acceptance region of rejection or region of acceptance. But when we are trying to work in the software, the

software try to compute the test statistics and they try to report the critical values using the concept of p values right. And it is possible that instead of using the concept of this C C star that is the critical region or acceptance region we can also use the p value concept to make the decision of the test right. So p value is essentially the estimated probability of rejecting the H0 that is the null hypothesis right. So based on that we try to make our decision rule and now when we are trying to test the hypothesis then we have two types of test which is one sided and two sided and in one sided we have left tailed or say right tailed test and two sided test means left tailed and right tailed both right.

So we try to compute the p value for a given statistics T x in the case of two sided hypothesis as probability of this absolute value of T is greater than T x under H0 when H0 is true right. And for one sided we try to compute the p value as the probability of T greater than equal to T x when H0 is true or T less than equal to T x when H0 is true for right hand side and left hand side hypothesis what is called as right tailed test or the left tailed test. I am not going to give you all these details, but I will try to show you in the examples so right. And the decision rule is extremely simple. The p value will be provided by the software and the level of significance that is the size of the critical region that is the value of probability alpha this is decided by the experimenter that can be 5 percent level of significance, 1 percent level of significance or whatever is more suitable for the given experiment.

For two-sided case: $P_{H_0}(|T| > t(x)) = p\text{-value}$

For one-sided case: $P_{H_0}(T \geq t(x)) = p\text{-value}$

$P_{H_0}(T \leq t(x)) = p\text{-value}$

Now so you have the value of alpha you have computed the value of p. Now if p value comes out to be smaller than the value of alpha that is the significance level then we reject the null hypothesis right and significance level is the probability of type 1 error alpha as we said earlier right. Now what are the steps when we try to conduct the test of hypothesis. So in order to test the null hypothesis we require a good statistical test which is developed by fixing the type 1 error and minimizing the type 2 error.

This is based on the Neyman Pearson lemma right. So these are the steps. So first we try to define the distribution assumption for the random variable of interest and specify them in terms of population parameter. For example that X1, X2, Xn they are coming from normal population with mu and sigma square. Based on that we can say that whether

mean mu is known or unknown or sigma square is known or unknown like this. Then we try to formulate the null hypothesis and alternative hypothesis right.

Then we have to fix the level of significance that is the type 1 error and then we try to develop a test which is equivalent to developing a test statistics right. Now this test statistics is constructed is computed on the basis of the given set of data right. And if my Tx which is a function of Tx1, x2, xn it is now computed on the basis of random sample and its value is calculated. And then we try to construct the critical region K for the statistics T that is the region where if the value of test statistics falls for a given set of data then we will decide that H0 is rejected right. And then we try to compute this value of Tx for the given set of data and then this test statistics is follows actually a probability distribution and that is the job of statistician to find out.

And then we try to find out the corresponding value of the test statistics based on the probability distribution from the table of say for example, if test statistics is follow a chi-square distribution then we try to find out the critical values from the tables of chi-square probability. If the test statistics is following the T distribution then we try to find out the value of the critical region or the critical values based on the probabilities of T distribution and so on. And then we try to compare these two values. The values which we have obtained on the given set of data and the value which we have obtained from the probabilities based on the distribution of the test statistic.

And the rule is like this. If the value of the Tx that is the value of test statistics based on the sample of data falls into the critical region then H0 is rejected and automatically the alternative hypothesis is then accepted. On the other hand, if the value of the statistics for the given set of data Tx fall outside the critical region then H0 is not rejected and H0 is accepted. And if you are working with the software then you have to look at the p value and if p value comes out to be smaller than alpha then H0 is rejected. The null hypothesis is rejected, right. And yeah there is a very good relationship between the decision based on test of hypothesis and the confidence interval, right.

And this is also used in making the right decisions. Sometime you will see in the software they will give you only the confidence interval and based on that you have to decide whether a particular type of test of hypothesis is getting accepted or not. Because the rule is very simple. If H0 is rejected at a significance level alpha then there exists a 101-alpha percent confidence interval which yields the same conclusion as the test, right. Suppose my null hypothesis is H0 theta equal to theta 0. Theta can be some parameter like as mu and we want to suppose test H0 mu is equal to say 20, right.

So theta 0 is some known value. And we try to create the confidence interval for theta. Now if the confidence interval of theta does not contain the value theta 0, theta 0 is your known value, right, then H0 is rejected. So this is how we try to take the same conclusion that we try to obtain on the basis of test of hypothesis using this confidence interval, right. So now we come to an end to this lecture and you can see here that I have given you some details that how are we going to conduct the test of hypothesis. And as I said that my objective was only to give you a brief overview, right.

There are many concepts which are related to the test of hypothesis. I have not given it here because it is not my objective here because I believe that you are aware of the test of hypothesis concept and I want to extend it to a multivariate case. So my objective was here to give you some knowledge so that me and you are at the same level and you know that what exactly you have to know in order to understand the lectures in the further course. So my request to you will be here that at least you try to brush up your concepts about the test of hypothesis and in the next lecture we are going to revise the test of hypothesis related to the mean in a univariate case in the case of normal distribution. And how to conduct it in our software that will be another issue. So I will try to start with the univariate case and then I will try to bring you to the multivariate case as well as when we have more than one random variable.

For example, it is testing of mean from a single population, from two population and then more than two population. Then conducting the test of hypothesis on a univariate random vector and then to a multivariate random vector. So that is why in order to prepare my background to explain you about the multivariate case it is important for me to first give you some information about the univariate case so that the understanding become easier for you. So you try to practice it and I will see you in the next lecture with more details till then goodbye. Thank you.