Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 07

Lecture – 32

Tests for Mean in Two Samples with Univariate Data

Hello friends, welcome to the course Multivariate Procedure with R. So, you can recall that in the last two lectures we have considered that test of hypothesis for the mean from a normal population normal mu sigma square under two situations when the population variance sigma square is known and when it is unknown. And we had considered there one sample test that means we had only one variable on which we are trying to obtain the data. In practice there are many situation where we want to compare. For example, I want to compare whether the population mean of two different population is the same or not on the basis of given sample of data. For example, if I say there is some medicine and suppose there are two medicines to control the body temperature.

Now both these medicines are given to a group of patient. Suppose I have group 1 to whom I have given the medicine number 1. I have group number 2 where I have given the patient the medicine number 2 to the patients. Now I want to see whether these two medicines they have a different effect or they have the same effect or also whether medicine 1 is giving a better outcome in the sense that it can control the body temperature for a it can maintain the or it can bring down the body temperature in the lower time etcetera.

So some questions can be increased about the comparison of the different type of body temperature in the two groups of the patient. So now suppose if I want to address these types of issue in simple way I can say I want to know whether medicine 1 is more effective or medicine 2 is more effective or whether medicine 1 and medicine 2 both are equally effective or they are they have different effects. So in this case we have two samples. So under these type of situation when we want to make a comparison using the test of hypothesis then how to do it? So I have here two condition that I have here two samples which are drawn from two different normal populations and these populations are differing with respect to the mean. Yes they can differ with respect to the variance also but in order to understand this concept over here I am considering that they are differing with respect to the mean only and they have got the same population variance.

Now this population variance which is common to both the population in the sense that they have got the same value that can be known or that can be unknown. So when it is known then in the case of one sample test we have used the Gauss test and we had used the command Gauss.test in R software and when this sigma square was unknown then we had computed it, we had estimated it on the basis of given sample of data and we had used the t.test command that was built in the base package. If you recall at that time I had told you that in these commands I had two options say X, Y where I had asked you to use Y equal to null because we were considering only the one sample test but now we are in a position where we are going to use both X and Y. So in this lecture we will consider both type of test of hypothesis when sigma square is known and when sigma square is unknown and we would like to compare the population means of two normal population. So let us begin our lecture and try to understand.

So now in this lecture we are going to consider that test for mean in the two samples with the univariate data. So first we try to consider the case when the population variances are known. So we have got here two populations which are differing with respect to the means and they have got variances but they are known to us. So we are interested in testing the hypothesis for the difference of mean. Difference of means means for example if I try to write H0: mu1 equal to mu2 then this can also be written as H1: mu1 minus mu2 is equal to 0 or in more general statement I can write in mu1 minus mu2 is equal to delta.

So and both these samples they have been drawn independently from the two population which are normal mu1 sigma1 square and normal mu2 sigma2 square. So now we have now these two samples from normal mu1 sigma1 square and normal mu2 sigma2 square and the first sample is of size n1 and it is drawn from a population whose mean is mu1 and variance is sigma1 square which is known and then we try to compute or estimate the sample mean which is here x1 bar. Then we try to consider the second sample of size n2 which is drawn from a population which has population mean mu2 and population variance sigma2 square which is also known and then we try to estimate mu2 by the sample mean x bar 2 and we wish to examine if two population mean mu1 and mu2 are the same or they are different. So now the null hypothesis is going to be H0: mu1 equal to mu2 and then we have here three possibilities to construct the alternative hypothesis. So one option is that I can consider the two sided hypothesis that is H1: mu1 is not equal to mu2 or other alternative is that I can consider one sided hypothesis H1: mu1 is greater than mu2 which is your here right tailed test or H1: mu1 is less than mu2 which is a left tailed test and this two sided it is a two tailed test.

- $H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$
- $H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 < \mu_2$

Now because population mean mu1 and mu2 are unknown then we try to use the statistics x1 bar minus x bar 2 to make some conclusion about mu1 minus mu2 right. So we are assuming here that sigma1 and sigma2 both are known, both the parent population are normal or I can assume that sample size n1 and n2 are large. If you try to see these assumption are similar to the test or say test that we have considered in the earlier lecture right. Now we know that sample mean x1 bar will be following a normal distribution with mean mu1 and variance sigma1 square upon n1. The second sample mean x bar 2 will be following a normal distribution with mean mu2 and variance sigma2 square upon n2.

So based on that I can write down the expected value of x1 bar minus x bar 2 is mu1 minus mu2 and variance of x1 bar minus x bar 2 is sigma1 square upon n1 plus sigma2 square upon n2 and the covariance between x bar 1 and x bar 2 will be 0 because they are drawn independently. Based on that we try to create here a test statistic, say here Zc. So if you try to see here this hypothesis H0: mu1 equal to mu2, I am trying to consider here as say H1, H0: mu1 minus mu2 is equal to 0 right. So now if you try to look at this statistic, say x1 bar minus x bar 2 minus 0, then this 0 is coming here because of this 0 right. In case if that is something else say delta, then this 0 will be converted to delta.

$$Z_c = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

So now I try to divide it by the standard deviation sigma1 square upon n1 plus sigma2 square upon n2 and square root. So this is the test statistics and when H0: is true, then this Zc follows the normal 0, 1 distribution. And so we try to use the probabilities from the normal 0, 1 distribution to compute the critical values and the p values right. So now if you want to implement this test in the R software, then as we have discussed earlier that we are going to use the command gauss.test which is available in the library compositions right, compositional data analysis.

So this has to be installed first and then this has to be uploaded. So this can be done by using the command install.packages and library compositions. And we have used it earlier in the one sample test case. Now the same command I am going to use it in the this two sample case also. So the command here will be Gauss.test. G here is in uppercase alphabets. Now you can see here this x is indicating the sample values from the population 1. This is sample 1 and y is the value which is in the sample 2 right.

Then mean here is 0. 0 is indicating this value, H0: mu1 minus mu2 is equal to 0. This 0 is this 0 here right. And standard deviation well it is here 1, but that you can give here any value right. And yeah on the other hand I can also say that if you are trying to consider H0: mu1 minus mu2 is equal to some delta then this 0 will be replaced by this delta value.

The alternative will remain the same as per your choice and the requirement as two sided less or greater as we have done in the earlier command right. So yeah this is the same thing where I just explain you. So now we try to consider an example and we try to implement it in the R software right. Suppose we have two samples which are about the gain in the weights in grams of fishes which are given two different types of diets say called as A and B. So, fishes are given two types of diets A and B with an objective to understand whether which of the diet is going to increase more weight right.

So the increase in the weights of these fishes they are recorded here as say in the diet A as like this in grams and in diet B they are recorded here as say in grams. And suppose the standard deviation is known to be here 2 right and we want to test if the two diets differ significantly as regards their effect on increase in the weight right. This standard deviation 2 it is actually the value of sigma1 square upon n1 plus sigma2 square upon n2 right. So I am just assuming it to be known. So now if you try to see I try to give this data into data vector xa and xb right.

Now I try to use here the command gauss.test and the first data set xa from here, second data set xb from here, mean is equal to 0 because H0: mu1 minus mu2 is equal to 0. Standard deviation is equal to 2 this is known this is given to us and alternative I am trying to take say two sided H0: or say H1: mu1 is not equal to mu2 and this is here the outcome you can see here. So you can see here the data here is used here like this and t here comes out to be here minus 2 this is the value of the test statistics and t value comes out to be here 0.009823 right. So now the rule here is that if p value is equal to 0.0098 and suppose you have chosen the level of significance alpha to be 5 percent level of significance then the rule here is reject H0: if p is less than alpha. So p here is 0.0098 and alpha is 0.05. So what do you think this is correct.

So in this case this H0: mu1 equal to mu2 is rejected that means gain in the weights from this diet A and diet B they are not equal or alternatively H1: mu1 is not equal to mu2 is accepted that is correct that both the diets are giving on an average difference increase in the gain in the weights right ok. So now this is here the screenshot and now I try to show you this thing in the R console also so that you can be more confident about it right. So first let me try to load this package although so now this here is library composition as I said I already had uploaded this thing in my computer now I try to give here the data here and we try to execute this command here. So you can see here this is here like this. Now if you want to change the alternative you can make it here less or if you want to make it here greater you can see here like this right.

So you can see here that the p values in the three alternative that is coming out to be different right ok. So now let me come back to my here slides and try to consider the other test. So now we move forward and our problem or the statement of problem is the same that we want to test the hypothesis about the equality of two means from two different populations which are the normal population whose means are different and their variances are not known. So up to now we have assumed that sigma1 and sigma2 both are known and for that we have used the statistics Zc which used to follow normal 0, 1 under H0. Now we are going to assume that both sigma1 and sigma2 or sigma1 square and sigma2 square they are unknown and in this case we are going to use the t distribution that means whatsoever be our statistic that is going to follow a t distribution.

Now this is the similar case if you try to recall that in the last lecture we have considered this setup in a single sample and we have used the command t.test. So in order to now extend it to two sample the same assumptions are going to continue, but then there is small variation. We are assuming that both populations are normal that means the samples are drawn from a normal population and we are assuming that both sigma1 and sigma2 or e to the power sigma1 square or and sigma2 square they are unknown, but they are equal. There can be other case also that they are not equal, but the test statistics for that case is different than what we are going to do here that is called as Fisher-Baron problem and that is handled in a different way and that is applicable when sigma1 square and sigma2 square they are unknown as well as unequal. And in the two sample case also if the sample size is large enough say greater than equal to 30 then we can approximate the t probabilities by the normal 0, 1 also and the reason is the same what we have discussed earlier in the case of one sample test.

So now we have here the same set of hypothesis that my null hypothesis is H0: mu1 equal to mu2 and my alternative can be mu1 not equal to mu2 mu1 is greater than mu2 or H1: mu1 is less than mu2. Now we are observing a sample of size m from normal mu1 sigma1 square and another sample of size small n, y1, y2, … yn from normal mu2 sigma2 square and both mu1 and mu2 are unknown. So, we try to estimate mu1 by say x bar and mu2 by y bar. So, I can use this x bar minus y bar to make conclusion about mu1 minus mu2 y because if you try to see this null hypothesis can also be written as H0: mu1 minus mu2 is equal to 0. So, now we have here both sigma1 and sigma2 both are unknown, but they are assumed to be equal.

So, now we try to create here the test statistics. So, we know that x bar will follow a normal distribution with mean mu1 and variance sigma1 square upon m. The sample mean y bar will follow a normal distribution with mean mu2 and variance sigma2 square upon n and expected value of x bar minus y bar will be equal to mu1 minus mu2 and now we try to compute or estimate the variance as two-way variance. How to do it? We try to estimate the variance sigma square by this expression 1 upon m plus n minus 2 and then inside the brackets summation xi minus x bar whole square plus summation yi minus y bar whole square. Now, based on that we try to construct here that test statistics which is here like this Tc.

So, once again if you try to see I had the null hypothesis mu1 minus mu2 equal to 0. So, we try to write down Tc here as x bar minus y bar minus 0, this 0 is coming from here right and divided by the standard error which is here like this right. So, this is my test statistics Tc which follows our T distribution with m plus n minus 2 degrees of freedom when H0: is true. So, in case if a sample is large then we can use the normal distribution to get the critical values or when we are trying to use the software then we can use the p values to get the to get the conclusion or the decision about the test of hypothesis. So, now let me try to take here the same example about the diet A and diet B which are given to faces and we want to test whether the effect of these two diets in increasing the weights of the faces is the same or different.

Alright, so this is the same example. The only difference is that earlier I had assumed the standard deviation to be known as 2, but now I do not know and I would like to estimate it on the basis of given sample of data. So, this is here the data on diet A and diet B which is stored in say two data vector x and xb, but now suppose the standard deviation is unknown, right. So, we will estimate it from the sample and now we want to test here the same hypothesis H0: mu1 equal to mu2 versus H1: mu1 is not equal to mu2. Now, if you try to see the command here, the command here is the same what we used had earlier is that is the same command.

Now the only difference here is this. Now I am giving here x as xa and data on y in the one sample case it was null, but now it is here given as here say xb that we already have understood in the earlier lecture and in the earlier command also. The alternative here is mu is equal to here 0, it is coming from here mu1 minus mu2 is equal to 0 and paired is equal to false. Why? It is not a paired test because we are getting the samples independently. So, samples are independent. If you recall if I in the earlier lecture I had explained you that under what type of condition the observations are going to be paired and then we try to use the paired t test right.

Now, the next option is variance is equal to 2. So, now you can see here I will assume that the two samples are coming from two different normal population which are differing with respect to only means mu1 and mu2, but their variances are unknown, but equal. So, now I have to give here true var equal that is variance equal is equal to true. The level of significant alpha here is suppose I take 5 percent to conf.level will be 1 minus alpha which is here 1 minus 0.05 which is 0.95.

Now, if you try to see here this is the outcome. So, this is the two sample t test data used here is xa and xb and this is here the value of t right and this is here the value of degrees of freedom m plus n minus 2 and this is here the p value. So, the p value hypothesis is 2. So, the two difference in the mean is not equal to 0 that is H1: mu1 minus mu2 is not equal to 0 and the 95 percent confidence interval here is obtained like this right and the sample mean of x and y they are obtained here like this 28 and 30 right. So, if you try to see that if I try to conduct it on the R console also this is a few shot here which I would like to show you now right.

So, if you try to see we already had stored the data on say xa and xb when we conducted the test when sigma1 square and sigma2 square are known xa is here like this xb here is like this and if you want to consider here this t test you can see here this is here the outcome. You can see here that it is the same outcome which I shown you on the R console. Now if you want to change here the oscillating hypothesis suppose if I want to take it here is less than type left tail test. So, in the left tail test it is coming out to be here like this the confidence interval is changing the p value is changing and if I want to make it here for greater than means H1 is mu1 greater than mu2. So, for that you can see here I simply have to change the alternative and you can see here that the t value will remain the same in all the cases the confidence interval is changing and the p value here is right.

So, now we come to an end to this lecture and you can see here that in this case what we have done whatever we have done in the last 2 lectures they have been extended from 1 sample to 2 samples, but if you try to see the way they have worked that is the same the methodology is the same. The only difference is coming because now we are using different test statistics when sigma1 square and sigma2 square are known or in general when the variances are known then the test statistics is following a normal 0 1 distribution and when the variances are unknown then it is following a t distribution. You have to be careful that when you are going for the case when variances are unknown then there are 2 possibilities that the 2 variances of the population they are equal or say unequal based on that you have to choose different statistics. Well, when they are unequal we try to that problem is actually called as Fisher-Béhrens problem, but now you see in the software you need not to worry you simply have to just change var.equal is equal to false and you are done and after that you have to simply look into the value of p and then you have to compare it with alpha. So, that is the advantage that when you are trying to conduct the test of hypothesis in a software then many things become easy, but it is important for you to understand that what is happening in the background that is why this theory is important.

So, now we have couple of jobs. First we consider one sample test then we consider then we extended it to sample test. Now would you like to extend it to a general setup when you want to or would you like to compare more than 2 population means? The next option will be you are considering here a univariate random variable, but would you like to consider it in a multivariate data? Then whatever you have done one sample test, two sample test etcetera would you like to extend it to a multivariate setup? Yes, that is our objective what we are going to achieve in the next lectures, but in order to understand them it is very important that by this time you have completely understood the philosophy behind the one sample test, two sample test under different type of conditions and what are the different methodologies as well as steps which are involved in the conduct of test of hypothesis.

So, for that I would request you, you try to create two data set yourself in one data set, you try to maintain lower variability and in other data set try to maintain higher variability and then try to see, try to compare them what happens or try to take the two data set which are of the similar natures, similar values and try to compare them, try to create different types of hypothesis, try to interpret them. The interpretation is very important, but one thing I would like to explain you here before I leave that all these tests of hypothesis they are based on certain rules of the statistical inference. Whatever Z statistics, T statistics either in one sample or two sample you have used, they are derived using the Neyman Pearson lemma.

And Neyman Pearson lemma gives us some methodology to derive such good test under the assumption that the null hypothesis has been constructed in such a way such that type I error is more serious than type II. Well, both are serious type I and type II both, but the way this Z test and T test whatever we have considered and whatever we are going to consider in the further lectures also, they are also constructed on the same assumption that the null hypothesis has been constructed in such a way such that type I error is considered to be more serious than the type II error. That means the consequences arising by the violation of the type I error are more serious in nature than the type II errors. Well, the reason is that it is not possible mathematically to minimize both the errors simultaneously. If you try to minimize one, other goes up and vice versa.

So that is why one is kept fixed, say alpha is kept fixed at 5%, 1% level and the second error is minimized in the best, to the best level. Yes, you can also ask me what if the type II error is more serious, then yes, you can derive different test statistics and then you can conduct the test of hypothesis, but definitely this is not the objective in this course, I am not going to tell you how to do this thing. So better is to use this existing test of hypothesis, test statistics and try to frame your null hypothesis in such a way such that type I errors comes out to be more serious than the type II error. So you try to practice it, try to take some data, try to expose them to the R software, try to understand the theory behind this test of hypothesis from the books and I will see you in the next lecture, till then goodbye. Thank you.