Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 07

Lecture – 31

Test and Confidence Interval for Mean in One Sample with Unknown Variance in Univariate Data

Hello friend, welcome to the course Multivariate Procedure with R. So, you can recall that in the last lecture we had discussed that testing of hypothesis about the mean in a normal population. When sigma square was known. Now, the alternative is what will happen if sigma square is unknown. And yeah, this is a very popular and means a feasible condition in real life. Because when we are trying to collect a data, then we have no idea that what is the population mean or the population variance.

So, in this lecture we are going to talk about the test of hypothesis for the mean when the sample is drawn by the normal population and sigma square that is the population variance is unknown to us. So, the methodology, the steps they are going to be the same what we have done earlier that we have to define our test statistics. Then we have to define its probability distribution so that I can find out the critical value from there and if you are using the software only then you have to depend on the p values. After this I will try to show you that how you can find out the confidence interval in such a situation and then we will understand that how are we going to conduct the test of hypothesis and confidence interval estimation in such a case when we have only one variable that is univariate case.

And it is important for you that you try to understand this lecture very carefully because whatever I am doing here in a univariate case later on I am going to extend it to a multivariate case. So, the procedure will remain the same the only thing will be the test statistic and the way it is represented they will change. So, this lecture is going to create a foundation for you to understand the testing of hypothesis in a multivariate case. So, let us begin our lecture and try to understand this procedure of test of hypothesis. So, now in

this lecture we are going to talk about the test of hypothesis and confidence interval estimation for the mean in one sample with unknown variance in univariate data.

That means, we have only one variable like as height or age. So, similar to the setup what we considered in the last lecture when the population variance sigma square was known we have a here a sample x1, x2, … xn from our normal population normal mu sigma square mu is the mean and sigma square is the variance. And we are going to conduct the test of hypothesis for the mean H0 mu is equal to mu 0 where mu 0 is known. Some constant value and this will be done by t test that is small t that is the name of the probability distribution if you recall we have done chi square sampling distribution t distribution f distribution. So, that is the same t.

So, in order to conduct such a test of hypothesis about the mean we have certain assumptions like sigma is unknown, population is normal and the sample size is smaller that is smaller than 30. And if you recall that I had shown you when we had talked about the t distribution that when the degrees of freedom are more than 30 then if this is my normal distribution and if I try to plot here a t distribution which is also symmetric both becomes almost the same. So, that is why when you have a small sample then we try to prefer this t test. And the difference will come only when you are trying to find out the critical value whether the critical values have to be found using the normal probability or the t probability. So, when n is greater than 30 then either you try to find out the probability from a normal 0, 1 or from the t distribution both will come out to be the same value.

So, that is why sometime in statistics we define the sample size to be smaller when n is equal to less than 30 is a small sample size and more than 30 is considered as last sample size. So, this concept has been changed because of big data and other things. So, now in order to conduct the test of hypothesis for S0 mu equal to mu0 we use here the statistics X bar minus mu divided by S by root n, where S is the standard error or S square is the sample variance which is computed from the sample data as 1 upon n minus 1 summation i goes from 1 to n, Xi minus X bar whole square. So, now the way we are going to conduct the test of hypothesis is that in the first system we try to fix the value of the level of significance at 5 percent say 0.05 or at 1 percent as 0.10 or so.

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

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

So, 1 percent, 5 percent, 10 percent whatever you want you can take it and now this Tc this statistics has got a t distribution with n minus 1 degrees of freedom under H0. So, I can write down here this is follow the Tn minus 1 degrees of freedom under H0. So, the critical values of this distribution can be obtained from the t table for a given with degrees of freedom and a significance level. And for last sample that is n is greater than or equal to 30, the T distribution can be approximated by normal 0, 1.

As I said that the both the probability curves of normal and T they become almost the same when n is greater than equal to 30. And if you are working on the software then we need to compute the p value which is computed by the software actually. So, this p value is computed depending on whether my alternative hypothesis is one sided or two sided. So, suppose small tc is the value of the t statistics that is computed on the basis of the given sample of data and this statistics capital T follows the t distribution with n minus 1 degrees of freedom. Then in the case of two sided t the p value is computed by this quantity twice the probability that T is greater than absolute value of T c for right tailed test like H 1 mu greater than mu 0 this p value is computed by probability T greater than T c.

And for left tailed test like H1 mu less than mu0 the p value is computed by probability that T is less than T c. And the rule is that we are going to reject the null hypothesis H0 in favour of H1 at alpha 100 percent level of significance if t value is smaller than alpha. Same that we have used in the earlier lecture also. Now, if you want to conduct such a test of hypothesis in the R software then the command here is t.test and it is available in the base package, right. You need not to install any other package like as in the case when the sigma square is known.

Now, the use of this command t.test is very much similar to the command that you use in the case of set test when sigma square was known, right. It is like this t.test and then after that there are various arguments. So, you have to give the data because we are considering here only the one sample test. So, data is going to be available only on one variable. So, I am writing here x otherwise the command here is like this t.test and then x and y here is actually NULL because as we discuss in the case of set test when sigma square is known that the same command is going to be used when we have more than 1 that is 2 samples.

So, in the 2 sample test also the same command is going to be used and there we will have to give the value of y which is here. Otherwise in the one sample case we are using it here as a null due to one sample. One sample means there is only one variable on which

we are going to obtain the data. Now, the alternative I have to choose whatsoever I am going to use whether 2 sided less or say greater. Mu is the equal to 0, this is actually the value from here mu is equal to mu0.

So, this is 0 actually is the value of mu0 right it can be anything then we have here paired. We also have a paired t test that is possible in the case of 2 sample that when we are trying to get 2 sets of observation on the same unit. For example, if a student is getting marks in a certain examination suppose here marks 1 and now that student is given some additional training and then once again the examination is conducted and the marks are obtained say for example, as marks 2. So, now if I say here there is only one student say student number 1 whose data points are collected marks 1 and marks 2. So, this is called paired data and in that case if I want to conduct the test of hypothesis then it is called as paired t test.

So, here in case of we are considering only the single variable. So, this paired is equal to false. Then we have another option here variance.equal is equal to false. So, this variance.equal is the situation in the 2 sample case when we say that the first sample data let us say x1, x2, … xn this is coming from some normal population with mean mu 1 and variance sigma 1 square and another sample y1, y2 say here y m this is coming from normal population with different mean mu2 and variance sigma2 square. So, this variance sigma1 square equal to sigma2 square can be true or false that means variance are equal or say unequal.

```
t.test(x, y = NULL, alternative =
c("two.sided", "less", "greater"), mu = 0,
paired = FALSE, var.equal = FALSE, conf.level
= 0.95, ...)
```

So, in the t test when we try to consider the paired t test and we assume that both the sigma 1 square and sigma 2 square are the same and in that case we try to give here the option variance.equal is equal to true, but since we have here only one sample. So, we are writing it here false and in case if a sigma 1 square is not equal to sigma 2 square then we have another test which is called as spatial variance test that is used, but anyway we are not going to consider here, but I wanted to explain you the interpretation of this command. So, that you can understand that under the given situation how are you going to use such options. Then confidence level it is here 0.95 actually you have to be careful that when I am when you have to give here the confidence level essentially we have to give the value of alpha as 1 minus alpha.

Alpha is the level of significance and 1 minus alpha is the confidence level. If you recall that when we constructed the confidence interval then we said there 100 1 minus alpha percent confidence interval is given by like this by those two limits. So, this is the same alpha here. So, you have to be careful here. So, if you are using here 5 percent level of significance then 1 minus alpha will be 1 minus 0.05 which is equal to here 0.95. So, this is how you are going to give all these things right. So, these are here the details of the same thing which I explained you right these are the same thing. So, you can read from the slide, but anyway I have explained you here.

Now, let us try to conduct this test of hypothesis on a given set of data. So, once again I am using here the same data set which I used in the case of when sigma square was assumed to be known, but here I am assuming it to be known. The reason why I am going to consider the same data set because I want that you can compare both the reasons that what really happened right. So, that data is about random sample of size 20 which is observed on the day temperature in a city and we assume that this temperature is following a normal use sigma square which is given where sigma square is unknown. And the value of the temperature in degrees calculate they are given here and they are recorded in a data vector temp like here this right.

So, now, I want to test here the hypothesis H0: mu equal to 40 versus H1: mu is not equal to 40. So, this is basically a two sided test. So, now, if you use the command t.test now you know what is here your x this is a temp that is the data vector. Why here is null because you are considering here a one sample test. Alternative here will be two sided why because here mu is not equal to 40 this is your here H1.

Now, mu here is 40. So, this mu equal to 40 is coming here by from here like this. And now paired equal to false because it is a one sample you do not have a paired data. Then we are assuming other option variance.equal is equal to false because you have only one sample and confidence level is 1 minus alpha which is 0.95. So, if you try to see here this is here the outcome that you will get and now we try to understand what is the meaning of different terms given in this outcome.

So, this is here data is equal to temp which we have used now. Next here is t is equal to 1.447 sincs. So, this is the value of x bar minus say mu0 which is here 40 by s by root of n, root of n here is a square root of 20. So, on the basis of given sample of data sample mean and sample standard error they are computed and its value is 1.4476. The degrees of freedom are actually n minus 1 which is 20 minus 1 equal to 19 and the t value here is

obtained as 0.164. And after that it is trying to tell us about the alternative hypothesis that true mean is not equal to 40. So, it is indicating that H1 mu is not equal to 40. Now, after this if you see it is giving you here 95 percent confidence interval.

So, if you remember in the last lecture I had told you that when we are using the package composition and use the command Gauss-Tosch test. In that case it was giving us only the result about the test of hypothesis and confidence interval was constructed separately, but I told you that it will not happen in the case of t test that is when sigma square is unknown. So, you can see here in the same outcome this 95 percent confidence interval is also given yeah. How to construct this confidence interval that I am going to explain you after this topic. And then after that so, this for the lower limit here is 39.12, this is the lower limit of confidence interval and 44.80616 is the upper limit of confidence interval and the value of sample mean which is here like this 41.965.

So, now you can see here that if you want to take here the proper decision then here p value here is 0.164 and alpha you have taken here as a 0.05. So, the rule here is reject H0 if p less than alpha. So, now you can take a proper decision because you can see here p is here 0.164 and is it less than 0.05? No, is it not true. So, that means H0 is accepted right and even if you try to understand as say H1 is rejected. So, H1 is mu0 equal to 40 is not accepted that means mu equal to 40 is correct value. So, this means that whatever differences you can see in the sample values you can see here all the values are pretty close to 40 and the average is around 41.965.

So, now you can see here you are trying to test that guess value is 40 and sample mean is giving you here 41.965 although there is some variation, but that variation is small and you can see that there is no significant difference between the population value and the value which you are trying to test right. There is not much difference between 40 degrees and 41 degrees because there is a lot of random variation also. And this random variation is there because you can see some value here is 32.8, some value here is 49.9 also, some value here is 55.62. So, there is a variation also. So, this is what we understand from the test of hypothesis. And this is here the screenshot of the same outcome that I have shown you right. So, let us try to first consider another type of test of hypothesis. I simply try to change the alternative nothing more than that.

So, the same set of data, but I now consider say less than type hypothesis this is one sided right. So, in that case everything will remain the same you can see here only the alternative is going to change as say less that is all. And you can see here this value here in the outcome t is equal to 1.4476 and degrees of freedom equal to 19 this will remain the same. In the earlier test you can see here they are thus they have got the same value.

The only thing here is earlier the p value was 0.164, but now this p value is changed to 0.918. So, that is obvious if you are trying to change the alternative hypothesis or the hypothesis which you want to test against the alternative hypothesis is changed. So, this p value is also expected to change. And you can see here in this case the confidence interval 95 percent confidence interval is also changed because this is from minus infinity to 44.3122 and, but the sample estimate remain here the same here right. So, this is the difference you can see here, but now you can use the same rule here that reject H0 if p value is less than alpha that is the level of significant p value here is 0.918 and alpha here is 0.05 and now so you can take a proper conclusion right.

And this is here the screenshot of the same outcome. Now I try to change my this alternative hypothesis and I make it here H1 mu greater than 40. So, all the commands everything remain the same only the alternative will change here as greater right. And you can see here that p value and degrees of freedom they will remain the same only p value will change again because alternative is now changed from two sided to less than to now to greater than right. And in this case the confidence interval will also change which is from now here 39.6178 to infinity right. And now you can see here that this p value here is 0.08 alpha here is 0.05 and your rule here is reject H0 if p value is less than alpha. So, you can use this and try to take a suitable command. Well all of them either you try to consider any of the hypothesis they will give you the same conclusion in a different way right.

So, now let me try to show you these things on the R console, but you can see here this is here the screenshot of the same outcome right. So, let me try to just copy this data and bring it to the R console. So, you can see here this is now here my data here temp and this command as I told you this t test this is available in the space package right. So, there is absolutely no issue.

So, you need not install any additional package. So, you can see here this is here like this and p value here is given 0.164 right. And if you simply try to change here only the value of the alternative that is from two dots either to say greater. So, you will see here you get here the values that t value they will remain the same degree of freedom will remain the same only p value will change.

The confidence interval will also change you can see here. And if you try to change the alternative hypothesis here see here less we can see here this t value degrees of freedom will remain the same in both the cases the p value is going to change and the confidence

interval will change. So, you can see here like this right. So, this is minus infinity to 44.3122. So, my objective was that I just wanted to show you all the three cases that how they work right.

There is no difference in the use of this thing, but I wanted to make you aware. So, that next time when we are trying to conduct the test of hypothesis then at that time even if I consider only one test of hypothesis or say one alternative only the remaining part you can do very easily right. So, now, the next topic I am taking here that how to construct the confidence interval in the same case that is the confidence interval for the mean and the sample in a normal population normal mu sigma square where mu is unknown and sigma square is unknown right. So, once again I assume that let x1, x2, … xn be a random sample from normal distribution mu sigma square where sigma square is now unknown. Let this data be here in this vector x1, x2, … xn and from here we try to find out the point estimates of mean and variance.

So, point estimate of mu will be x bar and point estimate of sigma square will be capital S square which are here given like this arithmetic mean and sample variance 1 upon n minus 1 summation i goes from 1 to n, xi minus x bar whole square. So, now, we know that x bar will follow a normal distribution with mean mu and variance sigma S square upon n and we have one more result here n minus 1 S square upon sigma S square will follow a chi square distribution with n minus 1 degrees of freedom right. So, now, we know that both these quantities both these random variables say here x bar and n minus 1 upon into S square upon sigma S square both are independently distributed. So, if you recall that we had discussed that when we want to construct a confidence interval we can use the vital quantity method and for that we have to define here a quantity whose probability distribution does not depend on any parameter.

So, this vital quantity here in this case is obtained here like this. This x bar minus mu divided by sigma by root n this follows a normal 0, 1 and the square root of this chi square random variable that is n minus 1 S square sigma square divided by its degrees of freedom n minus 1 right. If you consider this quantity which will come out to be like this square root of n x bar minus mu upon S, this quantity will follow a t distribution with n minus 1 degrees of freedom right. I am not asking you to have to understand the proof or derivation of the result at the moment well this is the job of the statistician and we teach these things in our BSc and MSc courses in statistics, but this result is important for you. Well, if you want you can look into the books and can understand how these things are derived they will help you in better understanding. So, you can see here both this parameters mu and sigma square they are involved in the normal 0, 1 case that is x bar minus mu upon sigma by root n, but and this vital quantity that g x bar theta mu sigma

square that is here square root of n, x bar minus mu upon S this is following a t distribution with n minus 1 degrees of freedom which is independent of any of the parameter mu or sigma square or any other unknown parameter.

So, that is why I can use this statistics to construct my confidence interval. So, now, I can say that a two sided confidence interval that is 101 minus alpha percent confidence interval can be obtained by writing the probability that square root of n x bar minus mu upon s lies between the two critical values minus t alpha by 2 at n minus 1 degrees of freedom and plus t alpha by 2 at n minus 1 degrees of freedom. So, that is probability is 1 minus alpha then I can solve this equation and I can write down very clearly here very easily actually that mu is lying between x bar minus t alpha by 2 n minus 1 degrees of freedom into S by root n and x bar plus t alpha by 2 with n minus 1 degrees of freedom into S by root n, right. Where this t alpha by 2 n minus 1 is the 100 alpha by 2 percent points on the t distribution with n minus 1 degrees of freedom that is t n minus 1. So, the 101 minus alpha percent confidence interval for this mu is obtained here as a the lower limit theta hat L is given here by this quantity x bar minus t alpha by 2 n minus 1 into S by root n and upper limit theta hat U which is here x bar plus t alpha by 2 n minus 1 S by root n, right.

So, now you have seen that this is how you can obtain the lower limit and upper limit of the confidence interval and now the 101 minus alpha percent confidence interval for mu comes out to be this x bar plus minus t alpha by 2 n minus 1 S by root n and it can be shown using the statistics tool that this is the shortest confidence interval. If you remember in the beginning we had talked about different types of concept that when the confidence interval is going to be say or what type of confidence interval is going to be good or bad. So, we had discussed that a shorter confidence interval is preferable, right. So, it is proved in statistics that if you try to obtain the confidence interval for mu from a normal population when sigma square is unknown then this confidence interval has the shortest length. Well, the length of the confidence interval is obtained by the upper limit minus lower limit that is theta hat L x minus theta hat U x, right.

So, in this case if you try to see here x bar plus t alpha by 2, n minus 1 S by root n minus x bar and this will become here plus t alpha by 2 n minus 1 S by square root of n. So, this x bar x bar get cancel out and the length of this confidence interval comes out to be t alpha by 2 n minus 1 S by root n, right. So, this is the length of the confidence interval, right, okay. Now, if you try to understand it graphically. So, if I try to take the t distribution of t with n minus 1 degrees of freedom and if I say that this area on the left hand side and this area on the right hand side is alpha by 2 and this area here is now 1 minus alpha and if you try to find out the ordinates here on the x axis then they will come

out to be minus t alpha by 2 n minus 1 and plus t alpha by 2 n minus 1 because t is a symmetric distribution.

Now, if you want to see the location of the confidence limits. So, this is here the lower limit of the confidence interval and this is here the upper limit of the confidence interval and this is here the area of this t distribution with n minus 1 degrees of freedom curve, right. So, this area is now here is the my confidence interval. So, this is how we graphically present it. So, I try to consider now here the same data where we have considered the temperature of on 20 days in degrees Celsius and now we would like to find out its confidence interval. So, in order to find out the confidence interval in the R software in this case that is confidence interval well for mu where sigma square is unknown and the population is drawn from the normal population in a univariate case the command here is conf.int, right.

This can be used along with the previous command t.test. So, although you have seen if you try to see here in this outcome this part is giving you here the 95 percent confidence interval you can see here, right. But if I want to find out this confidence interval here separately that means, only the confidence interval then how are we going to do it that is what I am trying to show you. Otherwise in the earlier command also t.test you can get this confidence interval. So, I am trying to store my data here as a in this beta vector here x and then I try to say here t.test and then x and then here conf.level that is the value of 1 minus alpha, alpha is here 0.05, 5 percent level of significance. So, confidence level become here 0.95 and now I have to write down here dollar and conf.int confidence interval that is the short form. And if you try to see here you will get here this type of outcome. So, you can see here this confidence interval is coming here, right. So, this is how you can obtain only the confidence interval and this is here the screenshot of the same outcome.

So, now let me show you this command in the R console also. So, I try to copy this data over here and this is here my data x or you can say temp also earlier we had used the terminology temp and if I try to use here this command here for the confidence interval you can see here this comes out to be here like this. And yeah in case if you try to change here the level of this confidence level suppose if I say 0.1 then you can see here this this interval is changing, right.

And if you want to make it here say here is only 0.5. So, what do we expect you can see here this is now here like this, right. So, this is how we try to construct the confidence interval in this case. So, now with this introduction with this we come to an end to this

lecture. And now you can see here in this lecture we have understood how are we going to conduct the test of hypothesis for the mean and how are we going to construct the confidence interval for the mean when the sample is coming from a normal population and the population variance sigma square is not known and how to get them in the R software that is also we have discussed. Now, as I told you during the lecture that in this lecture and in the last lecture that is the test of hypothesis for mu when sigma square is known and unknown along with it is a confidence interval we are going to extend it to a two sample case also.

In both that lectures you have seen that we have considered only one variable one cut of data either x or temp. But now if you want to consider the comparison of the mean of two population then the same command in R is going to be used with some modification. I have given you some idea in this lecture that how are we going to use the command t.test. So, in the next lecture we are going to consider the two sample test where I am going to use the same command.

So, I would request you that first of all you try to understand the t.test. So, the t.test command and gauss.test command what are they trying to do and how their interpretation is going to come when it come to the outcome of the software. The same thing will be repeated there itself and along with it I will say that try to take a data set try to understand these concept that what are you really trying to do and try to conduct the test of hypothesis yourself manually as well as in the software. And please do not stop here but try to understand the type of data what you are trying to take and the type of question which you are trying to answer is this matching. This I had shown you in the example that when I try to consider the temperature of the cities, I can see that most of the temperature are lying between 35 and say 50 degrees. So, if I am expecting that the average value should be close to 40 then it is not a bad option.

Well, there is going to be some difference because the values are varying. So, I am sure that even if you try testing H0: mu equal to 41 or mu equal to 42 then try to see what happens. Is your conclusion changing? And that is the same question which I ask you when I introduce that test of hypothesis that if I ask you that what is my age and I make a sentence my age is 20 years or I make a sentence that my age is 55 years how are you going to compare it. You will have certain information in your mind and based on that you will try to accept my statement more confidently that my age is 55 years then my age is 20 years. So, that is the same thing is happening here now I have shown you. But it is your turn that you try to practice it, try to take some more data, try to conduct it in the R software, try to understand what the software is trying to say.

Software cannot speak, software can give you the information it is your job or you have to understand, you have to learn how to understand what my this software outcome is going to inform me or what my software is trying to inform me. And if I understand it correctly I can interpret it better and I can give a correct statistical conclusion. So, you try to practice it and I will see you in the next lecture till then goodbye. Thank you.