Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 07

Lecture – 30

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

Hello friends, welcome to the course multivariate procedure with R. So you can recall that in the last lecture we have understood that what is the concept of test of hypothesis which is related to the testing the value of some parameter on the basis of given sample of data. Now from this lecture we are going to consider the case of normal distribution which has two parameters mu and sigma square and primarily we are more interested in developing a test of hypothesis for the mean. Now this can be done under two conditions when the variance sigma square is known or unknown. So in this lecture we are going to consider the case when sigma square is known and in the next lecture we will discuss the case when sigma square is unknown which is also estimated on the basis of given sample of data. Finally we will move forward and I will try to develop the similar test of hypothesis for the two samples as well as for multivariate case.

Well when I am saying here that we are going to develop the test of hypothesis for the mean in the case of normal distribution it does not mean at all that the test of hypothesis for other parameters from other type of probability density functions and probability mass function cannot be done. But as I have said several times our job here is not to learn the test of hypothesis but I want to give you here an exposure or some knowledge about those tools which I am going to use in the future lectures. So in this lecture we are going to consider the univariate random variable and we will try to construct the test of hypothesis when sigma square is known. Also we had understood in the last lecture that there is a close connection between this confidence interval estimation and the decisions on test of hypothesis.

And earlier we had talked about what are the confidence interval estimation. So in this lecture we will also consider the construction of confidence interval estimation under the same setup that is confidence interval for the mean when sigma square is known. And then I will try to show you that how you can compute these things in the R software. So let us begin our lecture. So now we are going to consider here the test and confidence interval for mean in one sample with known variance in univariate data.

So I already have explained you the concept of test of hypothesis and we are going to now work on it. So as we discussed that we need some assumption for every test of hypothesis. So when we are trying to test or develop a test of hypothesis for the mean when sigma is known then it is popularly called as Z test or Gauss test, GAUSS. And in order to execute this test or use this test there are certain assumption which are known to us. For example, the null hypothesis is going to be specified by the statement H0: mu is equal to mu0 where this mu0 is some known value pre specified value.

Sigma is known to us from some outside sample sources. For example, someone can tell me okay in this case the value of sigma square is supposed to be 25. We assume that the population is normal that means the sample which we are drawing from here belongs to a normal population. Or the second option is this if we do not know about the normality or in case if the sample size is here large which is more than 30. The reason for it is that when we try to consider the case when sigma square is unknown and is estimated from the sample then in that case the decision rule for both the case becomes almost the same.

So if you want to test this hypothesis H0 mu equal to mu 0 then the test statistics is like this. Zc is equal to X bar minus mu upon sigma by root n. So X bar is the sample mean which is obtained from the given set of data and mu will be coming from here mu is equal to mu 0 and sigma here is known and small n here is the sample size right. You have to be careful in the further lectures that where n is going to indicate the degrees of freedom. But here in this case this n is the sample size.

$$Z_c = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

And now we have to fix the level of significance that is the value of alpha say 0.05 or 0.10. And now the critical values that whether we are going to accept or reject the hypothesis on the computed value of Zc that is what we have to decide on the basis of distribution of test statistics Zc which has got a normal distribution with mean 0 and variance 1 when H0 is true. So that is why these critical values for this test they are obtained from the probabilities of normal 0, 1 right.

So I can say in general that Zc follows normal 0, 1 under H0 right. That is what I told you in the last lecture that we need to find out the probability distribution of that test statistics which is here Zc right. So now how are we going to conduct the test of hypothesis? We have to state our null and alternative hypothesis H0 and H1 then we need to compute the value of the test statistics Zc on the basis or given set of data. And then we have to obtain the critical value for a fixed value of alpha according to the hypothesis whether it is right tail, left tail or two tailed test. Then we have to compare the computed value of Zc with the critical value and then we have to make the decision accordingly.

Alternatively if you want to use the p value also then you can use the p value to make the decision about the acceptance or rejection of the null hypothesis right. So how to compute this p value? We have discussed in the last lecture just for your quick revision that here in this case Zc is the value of the test statistics which is computed on the basis of given sample of data and this follows the normal 0, 1. So the p value is given by this following probabilities. When we are trying to consider a two tailed test then it is computed twice the probability of Z greater than absolute value of Zc. For right tailed test it is computed as probability of Z greater than Zc and for left tailed test it is computed by here probability Z less than Zc.

And in case if you want to use it then the decision is like this that H0 is rejected in favour of H1 at 100 alpha percent level of significance if p value is less than alpha. And this p value is the smallest level of significance at which H0 would be rejected. So if you try to see when you are trying to work for the test of hypothesis that becomes very simple right. So now we see how to conduct this test of hypothesis in the R software. So this test is available in the R software in a package composition compositions which means compositional data analysis.

So this package is to be installed first using the command install.packages and then you have to upload the library compositions. Now this test is available by the command here gauss.test, g-a-u-s-s dot t-e-s-t but you have to be careful that g here is upper case right. So from this test we can have one sample test, two sample test with equal mean of normal variates and known variants.

What does this mean? Initially I am going to consider here only the one sample test but later on I will extend it to two sample test right. So in that case you will see the same package, same command is going to work there. Only small modification will be needed. So the command to conduct this test here is like this gauss.test and then you have to see here x which is the data on the on which we want to conduct the test of hypothesis x.

Why here is actually null because as I said this the same command can be used to test the equality of the means from the two population which is the two sample test. But here we are using it only in the case of one sample test. So that is why we are writing here y is equal to capital NULL. So if you remember once in the lecture we had discussed the concept of NULL that null is something which does not exist. Now we have to give here the value of mean M-e-a-n then we have to give here the value of standard deviation which is the value of sigma right.

And then now this alternative hypothesis can be of three type, two sided, less than, greater than. So that has to be indicated here by the option alternative is equal to two sided, less or greater and you have to choose only one whatever you are going. All these values are given here so that you can understand that there are three possible values and then you have to choose one value right. So you have to choose one option depending on your requirement right. So as I said that x is a numeric vector providing the first data set, y is the optional second data set which will be used when we try to conduct the two sample test of hypothesis mean is the mean to compare with as the standard deviation which is known and alternative is that the nature of the alternative hypothesis.

So now let me try to show you this with an example. Suppose we have a random sample of size 20 on the day temperature in a kitty and we have collected 20 observation and suppose the temperature in the population follows a normal distribution whose mean mu is unknown, but sigma square is 36 it is known. So this sample values of temperature in degree Celsius they are reported here like this and based on that we estimate the mu hat as sample mu that is 41.97. So if you try to see here this is the maximum likelihood estimate of mean right.

So on an average I can say that the temperature in the kitty at this time is 41.97 or close to 42 degrees Celsius right. This is what we mean. Now in case if I try to conduct the test of hypothesis based on this package composition. So I try to install the package, I try to upload the package using the command library, then I try to create here a data vector where I try to give all this data as a data vector temp right.

You will see that I will be using this example and the same data set in different lectures so that you can compare that how the things are changing right. Now if I want to test that the null hypothesis is mu equal to 30 versus alternative hypothesis mu is not equal to 30. What does this mean? I am trying to make a statement here that the value of the mean in the population is 30 degrees Celsius and my alternative statement against which we want to give the null hypothesis is that the mu is not equal to 30 that means the population temperature is not equal to 30 degrees Celsius. So I try to give here the data using the command Gauss.test x is equal to temp. Now y equal to here NULL, mean is equal to here 30. This is coming from here mu 0, 30 and as the you can see here this variance was given to be here 36 you can see here so this sigma becomes here 6, sigma square is here 36. Now alternative here is two-sided how? If you try to see here the structure of mu not equal to 30 from there I am trying to say that it is two-sided test of hypothesis and this is here the outcome. So you can see here this is here t, so this is the value of here Zc. Mean here is this equal to 30, this is the value of here mu 0, as v here equal to 6 this is the value of here sigma and now you can see here p value is coming out to be like this 1 into 10 power of here minus 6 which is very small, very close to 0 and the alternative hypothesis is here two-sided.

So now you can see here your rule was H0 is rejected if p value is less than alpha. So if you try to choose here alpha is equal to 0.05 which is 5 percent level of significance and your p value here is 1 into 10 power of here minus 6 what do you observe? That p value is much lower than the value of alpha, so I can say that here H0 is rejected. What does this mean? H0 is rejected means H0 is mu equal to 30 is rejected that means mu is not equal to 30 and if you try to see this also makes sense. If you try to see here that you had obtained the sample mean to be 41.97 that means the data is saying that the average temperature is close to 42 degrees and we are saying that the temperature here is mu equal to 30. So definitely there is a huge difference between the two value and that is the same thing it is implied by saying that H0 is rejected. So H0 is rejected means H1 is accepted. H1 is accepted mean, mean is not equal to 30 and that makes sense. Now it is not telling you that what is the correct value of mean.

Do you remember one thing? Test of hypothesis cannot tell you that what is the correct value of mean. It can only tell you whether the value you have chosen is correct or not. So now we try to change the alternative hypothesis and other things remain the same. H0 will remain as mu equal to 30 and H1 is mu less than 30. So the same data set I am going to use here, so the command here we will remain the same x value, y value, mean, SD they will remain the same only alternative is going to be changed as less.

So yeah I am just trying to show you that if your null hypothesis remain the same and alternative hypothesis is changed then how are you going to control it. So now if you try to conduct this test, the software outcome has this type of value. So you can see here this is data here is the same temp, t value will remain the same as earlier mean sd that is given. Now p value here is 1. So you can see here as you try to change the alternative hypothesis this p value is changing.

If you try to see here p is equal to here 1 and alpha suppose if you try to take here 0.05, so your rule was reject H0 if p is less than alpha. So in this case you can see here p value here is 1 and alpha is here 0.05. So you can see here now the decision is changed.

So and this is here the screenshot of the same outcome. So let me try to show you these things on the software so that you get here more confidence. So let me try to just copy this data over here and upload the library command. So composition command, so you can see here now here we have the temperature is here like this and now if I try to use say here this command here gauss-gauss test on the temperature data it is coming out here like this. You can see here this is the same value which is changed, right.

And if you try to use here the same data set but if you try to say here use here you can see here in this slide there were two options. I will try to show you here both this option less and greater, right. You can see here this is here less and this is here greater. So if I try to use here the command here greater you can see here this p value is here like this and if I try to change here this here less so then you can see here the p value here is now changed. So depending on your alternative hypothesis this p value is going to change and yeah in the first case where I am trying to take two-sided alternative that means H0 is mu equal to mu 0 and H1 is mu is not equal to mu 0.

Whereas when I am trying to take it here greater than so it is alternative here is mu greater than mu 0 and when I try to say here less then the alternative is mu less than mu 0, right. So this is how we can conduct such hypothesis or we can test such a hypothesis, right. Now I come to the confidence interval estimation under the same setup. So now we have got here a random sample X1, X2, Xn from a normal population with a known mean mu and known variance sigma0 square. So we have already used the point estimator which is this sample mean capital X bar to estimate the mu and now we want to construct the interval estimate, right.

This X bar equal to 1 upon n summation xi was the maximum likelihood estimate of mu. So now I am not going into that much detail but there is a central limit theorem which tells us that X bar follows distribution normal which is the mean mu and variance sigma square upon n. So here the variance is sigma0 square so X bar will follow a normal distribution with mean mu and variance sigma square upon n. Well this is the job of the statistician who are working in theory to derive such distribution and based on that if I try to create here a statistic square root of n X bar minus mu upon sigma0 then it will follow a normal distribution with mean 0 and variance 1, right. Well you need not to worry those who are not from the statistics background that is the job of the statistician to find out the probability distribution of such statistics.

So but now I know that because it is following a normal distribution so what I can do here that I would like to see for example if this is my here normal so I know that there will be some values on the left hand side and some value on the right hand side of this mean 0. So this value on the left hand side can be minus z alpha by 2 then the value on the right hand side at the same distance that will be plus z alpha by 2, right. So now I want to find out the probability that my here this statistics Z square root of n X bar minus mu upon sigma0 what is the probability that it will lie between minus z alpha by 2 and plus z alpha by 2 and I fix this probability as 1 minus alpha. What is this probability here? I am trying to take here this area as a alpha by 2, this area as a here alpha by 2, now this area will become here 1 minus alpha. So now I am trying to write down this equation that what will be the value of mu such that the probability that square root of n X bar minus mu upon sigma0 is lying between minus z alpha by 2 and plus z alpha by 2 whose probability is 1 minus alpha, right.

And if you simply try to solve this equation this will come out to be like this that mu should lie between X bar minus z alpha by 2 sigma0 upon root n and X bar plus z alpha by 2 sigma0 by root n, right. So yeah as I told you that square alpha by 2 is the 100 alpha by 2 percent points on the normal 0, 1 distribution or in terms of quantiles this is the alpha by 2th quantile on the normal 0, 1. So now I can say here that this is the lower limit of the confidence interval and this X bar plus square root alpha by 2 sigma 0 by root n this is the upper limit of confidence interval. And if you try to recall that when we introduce the concept of confidence interval estimation while introducing the estimation, right or statistical inference then we had indicated this lower limit to be here theta hat L and this upper limit was indicated here as the theta hat u. So now using the same terminology I can write down here that 101 minus alpha percent confidence interval for mu is this interval theta hat L theta hat U based on the sample X which is here like this X bar minus square alpha by 2 sigma0 by root n and X bar plus square alpha by 2 sigma0 by root n.

So you can see here if you get here some sample values so if you try to see here this is depending on X bar this is depending on z alpha by 2 this is depending upon sigma 0 and this is depending on here n. X bar this can be computed on the basis of given sample of data this z alpha by 2 is known from the table sigma 0 is already known and n is also known this is the sample size. So you can compute this interval and then you can see that with for example if I say alpha is equal to 0.05 then I can say that there is a 95 percent confidence interval for mu is that mu is lying between these two values. So that is the same if you try to recall once I take an example that when you want to compute the time taken from your home to your college then there are two option that you try to collect some data that for certain number of base you record the values of time taken and then you try to take the sample mean and then sample mean will come out to be some value that will be the point estimator and if you suppose sample mean comes out to be suppose you had 20 minutes then it is a point estimate but if you want to know then other alternative is that okay it will that you will take between 15 minutes to 25 minutes this was the confidence interval.

So now I have shown you here how are you going to construct this confidence interval based on the given sample of data. So this is the confidence interval for mean whose confidence coefficient is 101 minus alpha percent. So this 101 minus alpha percent this is called as confidence coefficient. So and if you try to understand the thing through this graphic so as I said suppose if you try to consider here the distribution of normal 01 so the mean is here 0 so this area here in the green colour which I am showing you this is the area alpha by 2 so corresponding value at on the x axis there will be z alpha by 2 on the right hand side and minus z alpha by 2 on the left hand side and on these two side there will be somewhere minus infinity and plus infinity. But when you are trying to compute the confidence interval now this will also follow on the normal contribution but now mean here is mu and this is lying between these two limits.

So the limit on the left hand side is x bar minus z alpha by 2 sigma 0 upon root n and this area here is alpha by 2 and on the right hand side the area is the same alpha by 2 but this limit is now here x bar plus z alpha by 2 sigma 0 by root n. So this is the basic idea and this area in say red this is 1 minus alpha. So this is the interpretation and the same interpretation will continue in all the cases whenever we want to do it. Now I would like to show you that how are you going to construct the confidence interval. Well in the case of confidence interval for mu when sigma square is known this is not available in the outcome of the R software but this will be available when in the software outcome when sigma square is known.

So in this case I am going to take the same example of the temperature where we have collected the data on say 20 days n equal to 20 and we have the data on the temperature and we are going to construct this confidence interval by writing our own program which is not very difficult. So if you try to see here now the temperature data is here now you can compute here if you try to see this confidence interval is depending on here mean and standard deviation. So simply have to write down here x bar which is obtained here by mean of temp and then you have to obtain the value of z alpha by 2 which is here norm 0.975 how if you try to see here this normal distribution this area here is 95 percent and so this area is here 5 by 2 percent and this area on the right hand side is 5 by 2 percent. So this is here means if you try to see here it is here I will use here a different color pen it is here 0 point means if you try to see here this is here 1 minus alpha by 2 and this area here is alpha by 2 like this.

So this comes out to be here 1 minus alpha by 2 which is here 0.975 and now sigma 0 is given to us and then square root of length of temperature which is the number of observation and if you and similarly you can compute here the upper limit by just by replacing minus by here plus sign that is x bar plus square alpha by 2 sigma by root n. So it is not difficult and if you try to compute it you can see here that it really had come like this. So now you can see here sigma 0 is here 6 as given and this Q norm can be obtained directly from the R software then square root of length of temperature that can be obtained directly from the data vector temp and then sigma 0 is given.

So this value comes out to be 39.33 and similarly if you try to compute the upper confidence limit this will be had come out to be like this just by replacing minus by plus sign right. This will come out to be 44.59. So I can say here that you can see here now if you try to consider this thing the point estimate was here 41.97 but if you try to see here that the interval estimate is between 39.33 to 44.59. So yeah close to the point estimate was close to 42. So now the interpretation is that on the basis of given sample of data I can say that the point estimator is 42 degrees and the confidence interval is 39.33 degrees to 44.59 degrees right and this is here the screenshot. I would try to show you this on the R console also so that you can have more confidence right. So I will just copy here this sigma 0 is equal to here 6 right temp data is already here because we have used it and now I try to write down here the lower limit which is here like this and upper limit if you want you have to simply make it here say here plus sign and you can see that it is here like this.

Okay so now we come to an end to this here lecture and you can see here that we have considered here a very simple test of hypothesis. My objective was to illustrate you how are we going to conduct such test of hypothesis. So we have considered the case when we want to test a hypothesis about the mean from a normal population and we want to test S0 mu is equal to mu 0 in a univariate case in a single sample when sigma square is known. Now similarly what will happen in the case when sigma square is unknown that is the actually more realistic case. So that we are going to consider in the next lecture but the process is going to be the same as we have seen that the statistics square root of an x bar minus mu upon sigma that is following normal distribution.

Similarly when sigma square is unknown and that is estimated from the sample then we would try to replace sigma by this estimated value and then we will try to see that the distribution of the statistics will become t distribution. What was t? If you try to recall sometime back you have done the sampling distribution chi square, t and F. So then we will see that how are we going to conduct the test of hypothesis but these steps and the way we have taken the decision for conducting the test of hypothesis and construction of the confidence interval that will remain the same. So it is very important for you that if you want to understand the further couple of lectures all the steps whatever I have explained you here they should be crystal clear in your mind. What are we doing? Why are we doing? What are the definitive steps? Yes some steps will be varying in comparison to what I told you in the last lecture because now since we are using the p value so that the test of hypothesis through simple in the case of this software becomes easier for us and it is pretty simple.

In comparison to you try to look into the z value from the table or similarly the other type of critical values from the table then can you try to compare it manually. So this is how we are going to proceed further. So my request is once again that you please try to look into this lecture try to revise it so that you are well prepared for the next lecture and I will see you in the next lecture till then goodbye.