Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 20
Hypothesis Testing: B

Welcome back my dear friends; a very good morning, good afternoon and good evening to all of you. And this is the Data Analysis and Decision Making - I course under the NPTEL, MOOC. And as you know this is a 12 week course of total number of hours being 30. So, there would be 60 lectures considering that each week we have 5 lectures and each lecture being for half an hour. And I am Raghu Nandan Sengupta from the IME department IIT, Kanpur.

So, if you remember we were discussing about hypothesis testing. So, in hypothesis testing you have some statement made by one entity, and you want to either verify that or refute that. So, when you want to basically verify that and refute that what you want to take is basically take a set of observations from the sample as a sample from the same population, and then based on your inference, your deduction you basically conclude that the statement being made by the other person is right or wrong.

Now, inherently as I mentioned that you always considered that the statement being made by the other person is false, hence you will basically try to have H naught framed in such a way that you will always try to deny that fact, considering that H A which is basically the alternate hypothesis you are trying in that to be true.

Now, in the other cases you should also remember that when you have some population para, basically something to do with the population parameter which you want to basically disprove or proof. And it can be other statement also. Now in the hypothesis testing if you remember the table which we discussed in the 19th class; so, the class just before this is the twentieth one as you can understand, there were 2 types of errors.

So, considering H naught to be true, actually being framed by nature or being given you prove that is true so; obviously, there is no error. Another case would be H A is true and you prove that H A is true so, there is no error. So, the error only comes on 2 occasions; that means, H naught to be true and you prove that H A is true; that means, you disprove

H naught case 1. Case 2 being H A is true, but from the other side and you give H naught to be true; that means, you disprove H A in that case there is the other error.

So, these errors actually remember are considered as an alpha and beta. And this alpha is intrinsically related to the level of confidence which was 1 minus alpha. Now if you go to the other example where we discussed about the bank and you being the manager of the bank, you are trying to give a loan the loan actual points, you will assign to people. And based on the points if it is more than 70 or 60 will give a loan and for people less than 60 or 70 will deny the loan. But there would be sets of people who should be given the loan, but you denied the loan that is opportunity cost lost. And there are other set of people who should not be given the loan, but you give the loan those are bad debts.

So, if you remember, I have drawn that slab that there are 2 distributions normal distributions with a certain mean values for both of them, the mean value for the people whom technically are not able to pay the loan are on the lower side, people who are technically able to pay the loan are on the right hand side. And the straight line which was basically the overall score based on which you will give a person a loan and not give a person a loan that was in between.

Now, trying to basically reduce alpha would increase beta or vice versa. So, I also mentioned that trying to basically reduce both at the same time is not possible, what you try to do is try to basically minimize the sum of them. But in technical terms when you solve the problem will keep beta fixed at a certain level, and then basically given the H naught to be true at a certain level of significance which is 1 minus alpha we will try to basically solve the problem.

Now, I also mentioned that in this hypothesis testing and basically the concept of interval estimations are basically follow up on the concept of point estimation. So, I will ask for forgiveness, but I am going back to the concept if you remember. For point estimation you use the maximum likelihood estimation method and the general methods of moments, I just discussed them conceptually I did not solve any problem and I only give the results.

So, once the point estimation problems have been found out, the and then you basically check using the concept of unbiasness and consistency, then you basically formulate the

interval estimation problem, depending on what is the interval and you basically find out the lower control limit and the upper control limit, and based on that you proceed. So, this lower control controlling with an upper control limit would be framed in such a way that they will depend on 1 minus alpha; which is the overall area covered in that pdf between the lower control and the upper control.

Also I mentioned that these values of upper control and lower control would differ depending on whether you want to find out something to do with the population mean or something to do with the population variance. If it is something to do with the population mean, there are 2 cases. Case 1: when you know the variance of the population, then you actually use the z distribution; obviously, there is no degrees of freedom for the z distribution. When you want to find out something with related to the population mean, but the variance of the population is unknown then you will use the t distribution. T distribution will be utilized considering there is degrees of freedom and you will lose one degrees of freedom.

Because you are trying to replace the population mean by the best estimate with the sample mean. So, when you utilize the standard error or the best estimate from the sample which is the best so called replacement for the population parameter which is the standard deviation of the variance, you divide by 1 minus one divided by n minus 1, and in place of mu you replace it x bar n. Why it is one divided by n minus 1? Because you lose one degrees of freedom, but as you are trying to utilize the set of observation which is x 1 to x n for the first time to find out something to do with the population mean using the best estimate which is the sample mean.

So, obviously, in the first case when you are using the z distribution the left control and the upper control and the right control; that means, the lower control and upper control limit would be a function of the mean value of the sample as well as z values whatever you take. And z values would; obviously, depend on the 1 minus alpha value which is alpha value. Then when we go to the finding out something to do with the population mean, given the standard deviation of the population not known.

Again you will find out the lower control limit and the upper control limit based on the fact that you have the sample mean the t distribution with n minus 1 degrees of freedom

and also s n. S n means where you divide by n minus 1. I am sorry I am repeating it time and again, but it will become clear to you as you solve the problem.

Then when once we came to the concept of the interval estimation; the interval estimation we utilize the concept for the population variance, you found out there were 2 cases. Case 1 when the population mean was known, in that case you use the chi square with n degrees of freedom; that means, you are not losing any degrees of freedom. And in the second case when you the population mean is not known, you utilize the chi square with n minus 1 degrees of freedom.

In the first case when you are using chi square with n degrees of freedom you will be using s dash. In the second case when you are using chi square with n minus 1 degrees of freedom you will be used using s without the dash. So now, in the third case when we come, you want to compare the differences of the 2 population mean.

So, obviously, in that case when you are trying to find out I am talking about the interval estimation case. And when you are trying to basically find out the difference of the population mean then; obviously, you will use the sample mean difference. But in that case you will first check up if both the population variances are unknown if the population variances are known you will basically use the z distribution. If the population standard deviation of the population variance is unknown, you will use the t distribution, but in this case remember that in the t distribution you lose one degree of freedom from the first population and one degree of freedom from the second population.

Now, when you come to, basically the case of trying to find out that ratios of the standard deviation of the population or ratios of the variances of the population, then you will use basically the f distribution. In the f distribution the degrees of freedom are technically m comma n, provided the population mean in both the cases are known. That is mu 1 and mu 2 are known. In case if mu 1 and mu 2 are not known, then you will basically use the f distribution, but with the degrees of freedom being m minus 1 and n minus 1. And basically solve the problem.

Now, when you try to utilize these concept which are talked in the case of hypothesis testing; remember the rules remain the same, but there are 3 cases for each of the

problems or the instances which I mentioned. Case 1, say for example, you want to basically find out something to do with the hypothesis testing based on the fact the H naught under H naught. I will not repeat H naught in each and every case. So, I will basically mention it for the first time now. And then basically only repeat the H A for each and every different instances.

Considering the H naught is the case where under H naught is the mean value is equal to mu naught. And there are 3 instances for this problem. Case 1 when H A is such where a mu a is less than mu naught, case 2 is in the case when mu naught is greater than mu a is greater than mu naught. And the third instance being when the case when mu a is not equal to mu naught. For all these cases technically the interval estimation problem was basically provided the population variance was known you basically use the z distribution and solved the problem.

So, in this case, you will have one set of problem the hypothesis testing; where the values you will reject H naught if and only if the values of the actual set of what if whatever the sample information is begin is given lies on the left hand side of the lower control limit. In the case when mu a is greater than mu naught, you will basically reject H naught if the values of the set of information which you get from the sample lies on the right hand side of the upper control limit.

And in the case when it is mu a is not equal to mu naught, you will basically have that it is depending on where the values of the sample information's are given, that mod of that difference between x bar n minus mu naught would basically be greater than and less than equal to the actual sample statistics which you found off.

Similarly, when you go to the case of trying to find out something to do with the with some happiness testing something to do with the concept of the population mean, provided the variance is not from the population is not known. You will use again the 3 instances: in one case, less than the lower control limit; next case, greater than the upper control limit and another case in between.

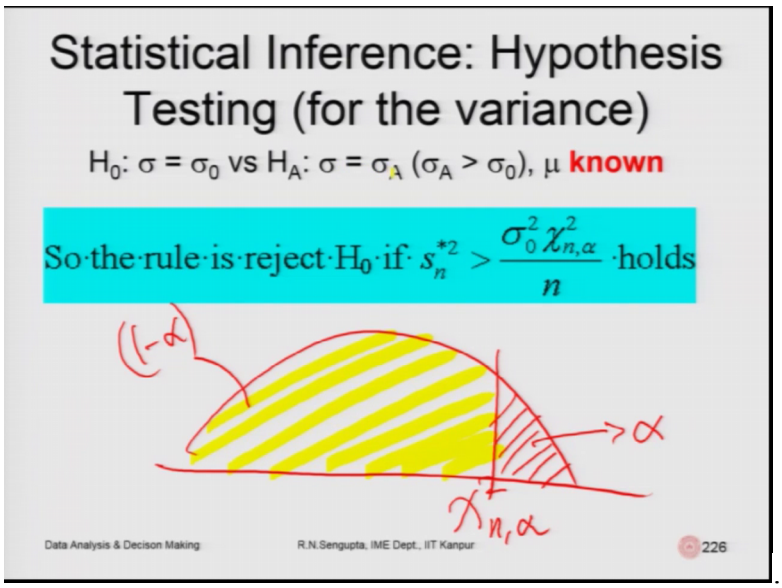
So, but in this case remember that in place of z distribution you will be using the t distribution with n minus 1 degrees of freedom. And obviously, remember another thing which I completely forgot to mention; that when you are trying to do the hypothesis

testing of the left than less than type greater than type the z value provided sigma square is known, would have a z value would have a suffix of alpha. But in the case when it is not equal to in that case z values on the left hand side and the left hand side would be z alpha by 2, because you will be taking the overall area in between the upper control and the lower control limit as equal to 1 minus alpha; which is the level of confidence.

Now, similarly when you come to the case of finding out the something to do with the population mean value provided the variance is not known, you will use the t distribution as I already have said few minutes back. But in this case the t distribution obviously; n minus 1 degrees of freedom remains true which is there is no problem in that. But in that case for the less than type you will use alpha; that means, the suffix in t n minus 1 will be alpha in the other case when we greater than time again it will be alpha the suffix and in the case when it is not a not equal to the suffix would not be alpha, but it will be alpha by 2. Because the overall area between the lower control limit and the upper control limit that is the lower value and the upper higher value should be 1 minus alpha.

So, we will be utilizing so, we discussed in this we will be we utilizing that time and again for the different types of parameter of the population you want to find out with depending on what set of information is given. So, whatever I missed is basically the recap in details about what you have been doing in the last 2 class which is 18th and the 19th one. And basically I will continue more in details discussing in the twentieth class also.

(Refer Slide Time: 13:48)



Now, considered that the statistical inference problem is the hypothesis testing for the variance only. And here if we consider the H naught is mu is equal to sigma is equal to sigma naught under H naught versus the case; where H A which is alternative hypothesis is says such that sigma is equal to sigma A and sigma is of less than sigma naught. So, in this case mu is known. So, if in this case is the mu is known. First I will draw the diagram; I will keep drawing the diagram for each and every rule such that it becomes clearer to you.

So, basically here the chi square; this is the value which you will have with n 1 minus alpha. So, in this case it will be chi square, n is the degrees of freedom because you are not losing any degrees of freedom, because if you remember this is when I am not going to highlight, but I am hovering my electronic pen. So, you see this s star; that means, you have not you are using the mean value of the population itself. So, there is no loss in the degrees of freedom.

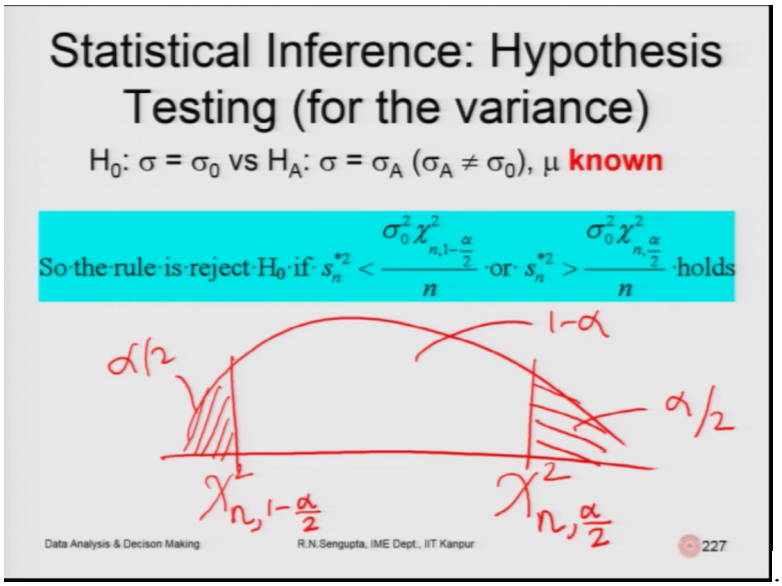
So, and this would be a 1 minus alpha based on the fact the area on to the right hand side basically would be given by 1 minus alpha. So, s dash square by n should be less than of sigma square naught under H naught into chi square n into 1 minus alpha by n. N is the degrees of freedom or the sample size. So, if this is true, then you will basically reject H naught.

Now, coming to the case where you have greater than type, so, I first draw the diagram and just mention that. So, in this case remember again the rule is s dash square is greater than sigma square naught into chi square n alpha by n. But now note down important thing it is alpha and 1 not 1 minus alpha. This is the nomenclature.

So, alpha being the fact, that the area still to be covered on to the right is alpha. In the case, when he had the z distribution, the z values were symmetric that is why they could be read written down as minus z alpha and plus z alpha. Similarly, for the t distribution as samples size increases, it becomes symmetric. Hence also this nomenclature of using in the suffix, 1 minus alpha or 1 minus alpha by 2 with respect to the case of alpha r 1 minus or alpha by 2, does not happen, because that is not relevant. But in this case it becomes relevant for the chi square and the f distribution.

So, in this case you will basically reject H naught if s dash square is greater; that means, on the right hand side. Now when we come and obviously, the hypothesis is H naught under H naught mu is equal to mu naught. And with sigma is equal to sigma naught with respect to the case that H A being sigma is equal to sigma A where sigma is greater than of sigma naught and mu is known, the mean value of the population is known.

(Refer Slide Time: 17:48)



Now, the final case for this under this is the statistical inference problem for the hypothesis testing. Under H naught sigma is equal to sigma naught, and under H A sigma

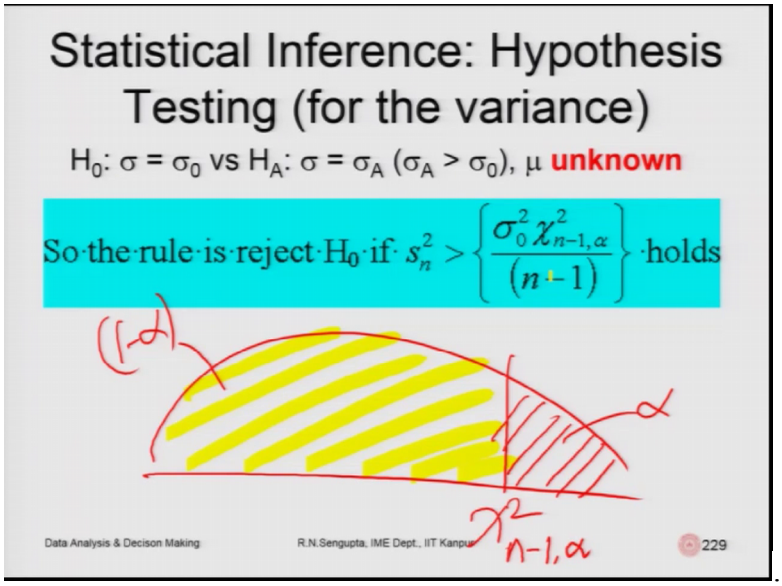
is equal to sigma A, but in this case sigma A is not equal to sigma naught, and in and the mu mean value of the population is known. So, let us basically draw the distribution if so, you have 2 values. So, this chi square first write down the chi square so; obviously, you are not losing any degrees of freedom because the population means value is known.

So, you can safely use n, you can safe to use n. Now comes the interesting part for the nomenclature thing. So, this area and this area should be alpha by 2. This area should be 1 minus alpha. Now what you will try to do is that the area for know under the nomenclature scheme the suffix would be 1 minus alpha by 2 that is the overall area which is still to be covered on the right. And this will be alpha by 2 corresponding to the fact that the area still to be covered on the right is alpha by 2.

So, if you note down it this becomes clear. So, in this case you reject H naught if s dash square; obviously, dash is true because mu is known is less than on the left hand side sigma square chi square n. And 1 minus alpha by 2 which is basically what we have basic written. So, it corroborates the fact that, then are the nomenclature and the way we are trying to basically give the rules acts actually absolutely match.

So, there is no confusion. So, this becomes 1 minus alpha by 2 as it is. And in other case it will be if it was greater than times because if you are you are taking the fact that is not equal to for the greater than time. S dash square is greater than chi square sigma square naught chi square n, and this is alpha by 2 as rightly noted down it is alpha by 2 divided by n. So, obviously, both the rules as well as the nomenclature are matching and there is absolutely you know no error no confusion here.

(Refer Slide Time: 20:15)



Now, let us come to the fact that we are trying to find out something to do with the variance, but the population mean is not known. So, if the population means not known, the next question would be coming that; obviously, we will be using the chi square, but in that case the degrees of freedom would be reduced by 1, because you are utilizing the sample set of observations x 1 to x n for the first time to find out something to do with the population mean by the best estimate with this is a sample mean point 1.

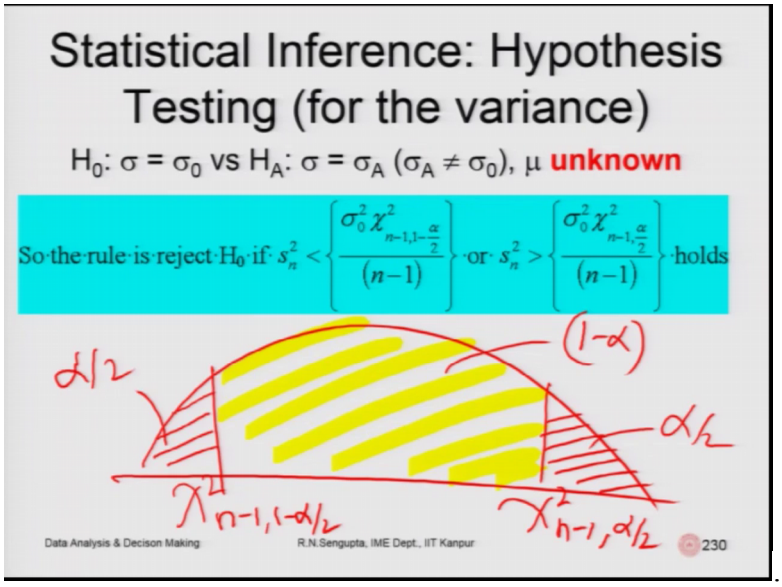
Point number 2, in the case we are using the sample mean; obviously, in that case s dash is not used, it is s because we have lost one degrees of freedom. So, s is the formula where you divide by n minus 1, and in place of mu you use x bar n. So, let me draw the distribution chi square the left part is chi square, this area is alpha, this area is 1 minus alpha. This would be chi square n minus 1, and 1 minus alpha because on the right hand side. So, I highlight it so, less than type so; obviously, it will reject. So, let us see this chi square n minus 1, 1 and small alpha is exactly here.

So; that means, you will reject H naught if s square by n, n means basically the sample size is less than sigma square naught under H naught chi square n minus 1, 1 minus alpha and divided by n minus 1 because you have lot lost one degrees of freedom. So, this is the hypothesis testing for the case when under H naught sigma is equal to sigma naught, under H A sigma is equal to sigma is equal to sigma A, but sigma is less than sigma naught and mu is unknown.

So, of the greater than type what it happens? The room the hypothesis testing is under is H naught, sigma is equal to sigma naught under H A sigma is equal to sigma A and, but sigma is greater than sigma naught and mu obviously is unknown. Chi square is obviously it will be n minus 1, and this will be alpha. This will be 1 minus alpha, and use the highlighter. So, if you consider this let us see the nomenclature which we are using and the rule they match.

Obviously we should come to the rule and match with our nomenclature. So, also vice versa whatever it is, but let us see it is matches. So, you will see the rule is to reject H naught if s square n is of greater than type. On to the right hand side of sigma square suffix not, chi square n minus 1 which is right. Alpha because the area on to the right hand side is alpha that is the nomenclature is matching divided by n minus 1; which is the due to the fact that you have lost one degrees of freedom.

(Refer Slide Time: 23:27)

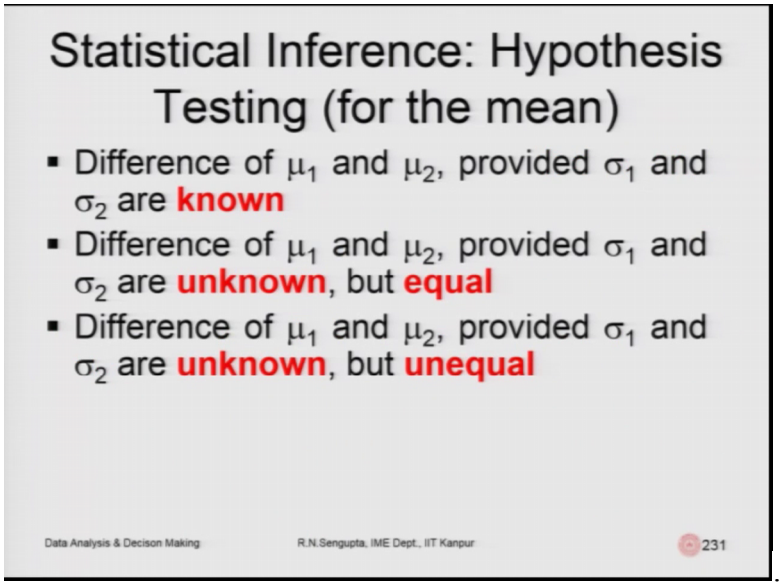


Finally, for the case so, I will first write the degrees of freedom. So, this is alpha by 2. So, this becomes alpha by 2. This would be 1 minus alpha by 2, because this is alpha by 2, this is 1 minus alpha. So, this is we are drawn let us see. So, on the hypothesis is H naught, mu is equal to mu naught now sigma is equal to sigma naught versus H A where sigma is equal to sigma A sigma is not equal to sigma naught and mu is unknown.

So, in this case it is on the left hand side or the right hand side. Let me highlight it has I have been doing sorry for that. So, if you are following on the same type of trying drawing we should follow it. So, in this case rule is rejected; if I am that means, reject H naught if s square is less than type, on the left hand side on the right hand side.

So, the first case is less then type this sigma square not, chi square n minus 1; obviously, and this is 1 minus alpha by 2 which is matches; divided by n minus 1 because you have lost one degrees of freedom. And or the other case is true where s square is greater than sigma so; that means, on the right hand side of sigma square not guy square n minus 1 by alpha by 2 which is also matching with the nomenclature and as the rule says divide by n minus 1.

(Refer Slide Time: 25:09)



Now, we will just mention I would not give the rule, but I was tried to solve it later on with the problems. So, difference if you want to find out the difference between mu 1 and mu 2 provide a sigma 1 and sigma 2 are known. So, what do you do? So, first pause let us think. So, if sigma 1 and sigma 2 something has to be found out; obviously, that distribution has to do with something true with z and t they cannot be anything else point 1.

Point number 2, let us find out what are the sets of information as given it is given as sigma 1 and sigma 2 unknown. So, if it is known so; obviously, you can rule out t

distribution so, it is z. So, in this case you will do to z distribution, if it is less than time, it will be minus at alpha, it is greater than type, it will be plus z alpha, not by 2. Because not by 2 would not becoming and if it is not equal to; then only minus at alpha by 2 and plus z alpha by 2 point 1.

Point number 2 the best estimates for the difference of mu 1 and mu 2 would be x bar n 1 minus x bar n 2 where n 1 and n 2 are the sample size which you are taking from the first population, and n 2 is basically the sample size we are taking from the second population. Hence x bar n 1 and x bar n 2 are the best estimate of mu 1 and mu 2.

Now, obviously, you will basically have a pooled variance, I will come to that formula pooled variance means basically the what will be the variance of both of population 1 population 2; where the standard deviation or the variance for population 1 and population 2 are given as sigma 1 or sigma 2 one square and sigma 2 or sigma 2 square.

Now and when it is not equal 2; obviously, I mentioned that I am again repeating it. Then in that case it will be z alpha by 2 on the left hand side and z alpha by 2 on the right hand side. When it is the difference of mu 1 and mu 2 provide a sigma 1 and sigma 2 are known, but equal in that case be rest assured that if there are unknown and if they are equal then, and then immediately the word unknown should strike. In that case z distribution only ruled out you will use the t distribution.

Now, in the t distribution when there was only one population you lost one degrees of freedom. But in this case as you have 2 populations you will do at least lose 2 degrees of freedom. So, t would be with the suffix n 1 plus n 2 minus 2 where n 1 and n 2 are the pop sample size from population 1 and population 2 that is point 1.

Point number 2, when you are trying to find on the difference of mu 1 and mu 2; obviously, their best estimate from the sample would be x bar n 1 and x bar n 2. So, where x bar n 1 and x bar n 2 are based on the fact that n 1 and n 2 are the set of samples which you are taking from population 1 and population 2. And obviously, the pooled variance would now be replaced by the corresponding standard errors from population 1 and population 2; that means, from the sample one and sample 2 and you will calculate accordingly.

Here also in the case when you have left hand side; that means, less than time it will be t I am just talking about the main points. It will be t suffix n 1 plus n 2 minus 2 and it would be given by alpha. Because of the less than time for the greater than type again it will be t suffix n 1 plus n 2 minus 2 alpha. And in the case when it is not equal 2 it will be given by t, on the left hand side t n 1 plus n 2 comma minus 2; obviously, comma alpha by 2 and similarly for the right hand side it will be t n 1 plus n 2 minus 2, and then in that case also it will be alpha by 2.

But alpha by alpha by 2 and for the z 1 if the t distribution would be would be as the asymmetry; hence the concept as we have used for the chi square and the f distribution would not be used; that means, 1 minus alpha by 2 and alpha by 2 those things would not be utilized. Then again when we have the difference of the mu 1 and mu 2 provided sigma 1 sigma 2 unknown, but on unequal the whole concept remains the same as second bullet point only in this in the sampled pooled variance or the standard error for the pooled sample with the formula would change. And you will utilize the same concept for all our calculation.

With this one I came this lecture I will end the twentieth lecture and continue more discussion about the chi square distributions from in problems and the f distribution problems. And then slowly go into the multiple regressions. So, these parts would be more applicable as you solve the problem. As I basically discuss very simple problems and if you can basically solve it to yourself. Both the assignments obviously assignments are from the grading point of view for this NPTEL MOOC course, but if you basically read the book and try to solve those things would be much clearer for you. And thank you very much for your attention, and have a nice day.

Thank you, bye.