Total Quality Management - II

Prof. Raghunandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

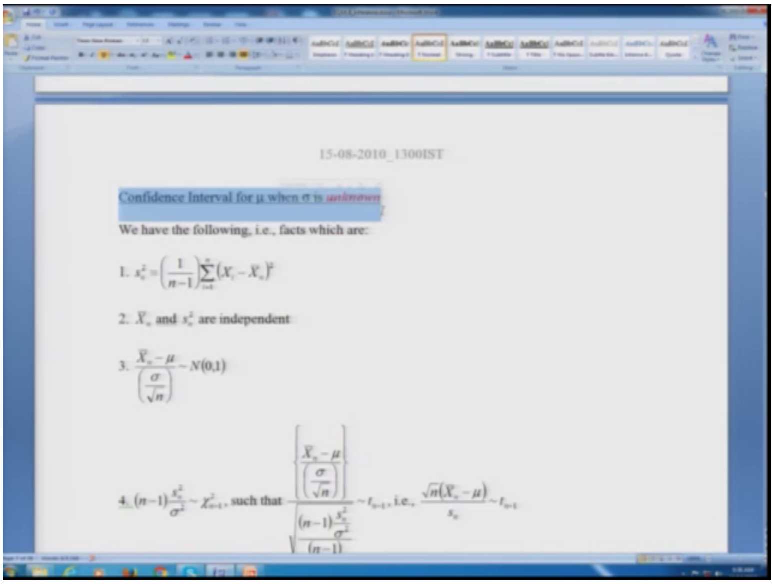
Lecture – 05
Part-2
Interval Estimation- II

A very good morning, good afternoon, and good evening my dear friends, welcome to the TQM - II lecture series which is another NPTEL mooc and as you know it is a 20 hours program. So, each week you do 5 classes each of half an hour duration and after each week you have assignments. So, it will be 8 assignments plus 1 and semester examination on in term examination or final examination whichever you name it. I am Raghunandan Sengupta from the IME department IIT Kanpur. So, we will be starting the fifth lecture and as you remember that we have discussed that all the things related to design of experiments and further related topics in TQM and other relevant areas.

So, you basically need to understand the concepts of the 4 main it distributions which is the normal 1. Obviously, we did discuss the discrete and the different type of continuous distribution, they would have implication for the hypothesis testing later on considering the central limit theorem to be true. I did mention central limit theorem very fleetingly, but it is essence without going to the prove it is essence and it is importance will come out later on and which do more and more discussion about that.

So do not worry about that we will do that and the main 4 distributions was normal and from that normal we had basically the t distribution the chi square and the F and if you remember I did this mention it and it I will be repeating it time and again that is anything to do with the mean value or the first moment for the particular distribution would be related to the z or the t and anything to related to the standard deviation of the variance which is the second moment would be related to the chi square and F and I also did mention that the concept of degrees of freedom that will become very apparent as you start going through the depth, which we I will try to cover in lecture 5 and hopefully if wrap it up in lecture 6. So, consider now the problem is like this initially and if you see the slides or the document file to word doc.

(Refer Slide Time: 02:40)



So, that would make it very clear to you in the initial part we had in the last lecture fourth 1 at the fag end, we had to do something to do with the mean and the standard deviation of the variance was known. Now we will basically consider something to do with the mean with the standard deviation and our various not known, it may be a repetition what we did in class in lecture 4 n part, but please bear with me. So, we would basically have a best estimate from the sample on the best proxy from the sample which we will try to basically be the best estimate considering un biasedness and consistency, those properties have to be checked which I did not mention unimportant, but I did not go through the proofs that is not required for this course.

So, you had basically s dash and s s dash is the case when the is the best estimate for the population variance using the standard error or the so called Variance of the Sample, considering the mean value is known. And in the other case when the mean value is not known you use s without the dash this is what we are considering, because if we read the top line it basically mentions where I am hovering my or I am highlighting it basically means confidence interval for mean considering the standard deviations are unknown. So, this confidence interval obviously, once you understand it will have a very simple understanding in the hypothesis testing.

So, we have the following facts which means s square which is the square of the standard error for the sample considering the mean values unknown, would be given by 1 by n minus 1 because minus 1 is coming due to the fact that you are losing 1 degree of freedom and then it will be summation and the sum would be of the squares of the values of the difference between the random variable and it is mean values mean value is the sample mean which is the best proxy for the population mean.

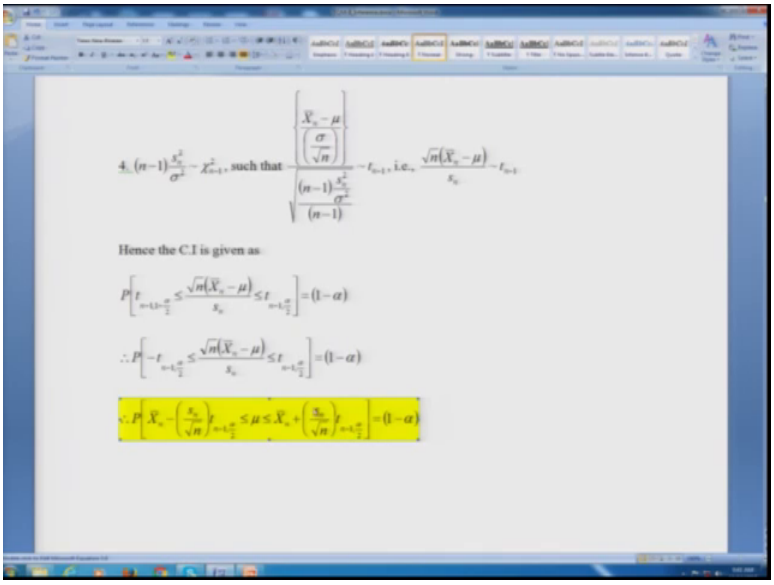
We also know for the normal case X bar which the sample mean and the s without the dash square are independent and obviously we will know that X bar n is normally distributed with mean mu and variance sigma square n by n. Hence it can be converted into a standard normal or z distribution. So, once this is done you can easily find out that you can get 2 type of distribution from them, 1 is basically the case of the chi square where I am not going to highlight I am just hovering the pen above.

So, electronic this electronic pen, so this is basically the chi square now you note down the degrees of freedom it has lost 1 degrees of freedom, because you are trying to find out the ratio with of the square of the standard error divided by sigma square for the population on the variance of the population. But this square which is in the numerator is basically the standard error holds whole square considering the mean values unknown. So, I will be repeating this time and again please bear with me and if you remember the t distribution with t distribution is basically the ratio off and in the third and the fourth lecture we did discuss with the ratio of our z distribution divided by the square root of a chi square divided by it is degrees of freedom.

So, now so the numerator part where again I will try to highlight. So, this part the numerator on this highlighted equation numerator is basically a standard normal and in the denominator which were the cursor is basically a chi square, but that chi square divided by it is degrees of freedom, but the chi square is already we know is of n minus 1 degrees of freedom; so obviously, we will get a t distribution with n minus 1 degrees of freedom which is very logical.

Now, given this you will basically try to find out the lower limit and upper limit, considering that within that bound if you remember it is discussed within that bound and the overall confidence that the upper limit and the lower limit would have the so called parameter of the population, within that is given by 1 minus alpha which is the confidence level and if you find out the confidence level we want to, so we have to basically formally did accordingly.

(Refer Slide Time: 06:59)



So, probability of the lower limit corresponding through the t distribution and the upper limit of the t distribution which is basically upper limit being t, now mark my words carefully n minus 1 is the degrees of freedom and comma alpha by 2 because, this is the area which you have to cover more on to if you are looking at the distribution from your end and this is t if you remember has a has a mean value which is 0 on to the left are the negative values on to the on your right are the positive values.

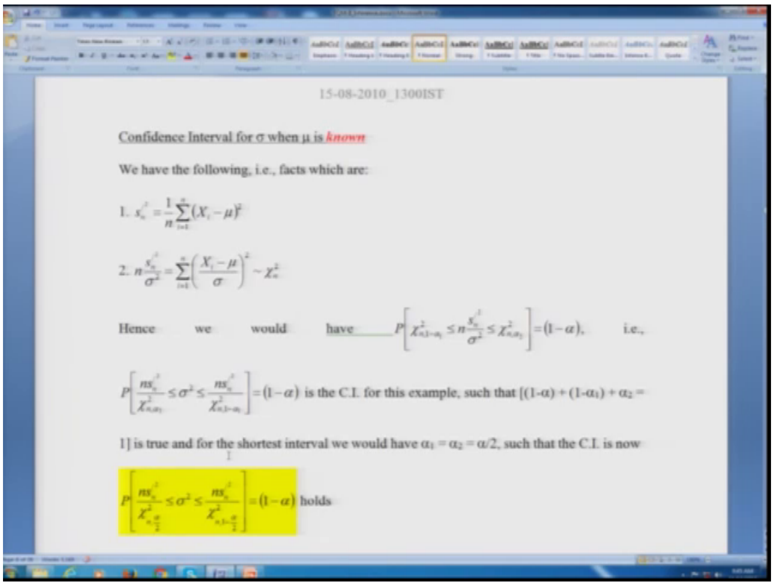
So, if you go subsequently along that X line which has the t values, so obviously they would come a certain upper limit where the t values live aside the degrees of freedom, we have already discussed that it will be alpha by 2 corresponding to the case that the area still to be covered is alpha by 2. Now if you come on to the left hand side left hand side from your side which is the lower limit it would be t again n minus 1 is very obvious, it will be 1 minus alpha by 2 which means the overall k area to be covered more further on to the right.

Now, you may be thinking that we I had been mentioning that the only distribution and as you know which is basically absolutely symmetric is the normal distribution, but t distribution if you remember if we increase the sample size n even for small sample size it is symmetric; hence this total confidence level is 1 minus alpha. So, the whole overall area left is alpha, so it is being equally divided on to the left hand side and the right hand side with values of alpha by 2. So, if you basically do some juggling very simple like taking the new numerator denominator the numerator and cross multiplying and all these things.

So, at the end of the day the result is where I am highlighting let me highlight is further on. So, the lower value would be X bar n minus s n divided by square root of n into t n minus 1 alpha by 2, because as it is symmetric hence 1 minus alpha by 2 and alpha by 2 would be of the same numeric values, but with the sign being therefore 1 minus alpha. So, hence this n n minus sign is coming where the cursor is now being hovered around. So, that would be the lower value and the upper value, obviously would be X bar n plus because that is on the right hand side from your side would be plus s n by square root of n t n minus 1 alpha by 2.

So, that would basically be the lower end I mean upper limit between that you will have the mean value and that the total confidence level is basically 1 minus alpha. Now if I want somebody wants to find out the difference between this the lower limit and the upper limit at the length of the confidence interval. Obviously it will, me the upper value minus the lower value. So, the basically X n bar X n bar cancels it will be twice of X s n divided by square root of n into t n minus 1 alpha by 2.

(Refer Slide Time: 10:01)



Now, consider now we had we have seen in these 2 examples last day and today, that they main focus was to do with something with the first moment which is the mean hence we use the z distribution to find out the conflicts develop the mean, provided the variance of the standard deviation and the population is known. In the second case we tried to find out just we finished few seconds back is basically to find out the confidence interval and for that we use the t distribution with 1 degrees of freedom being reduced which is basically going from n to n minus 1, for the case when something to do was to find out the confidence interval of the mean value of the population provided the standard deviation or the variance of the population is unknown.

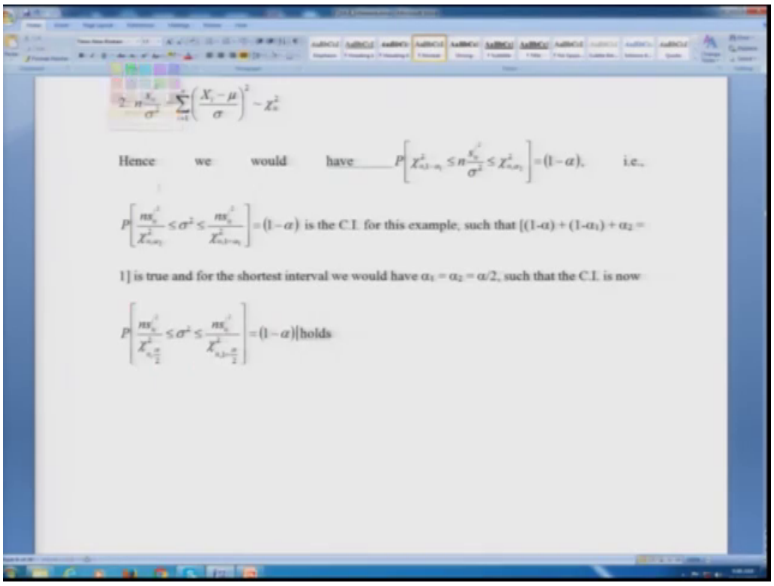
Now, we switch our overall discussion to do something with the quant with confidence interval study and the study of the of the interval estimation to do with the standard deviation of the population which is the second moment. Now in this case again you will have 2 scenarios scenario 1 would be when the population mean is known and scenario 2 is the population mean value is not known, so let us now proceed. So, as shown the confidence interval for sigma when mu values is known, so obviously we will use with a little bit logic we can understand we will use s dash square because, in n s dash we do not lose 1 degrees of freedom and because, the mean value is the population mean so s dash is given.

Now when we try to find out the ratio of s dash square by sigma square as we did in the last case, but in that case the ratio in the numerator was s without the dash when you do find try to find out the ratios it becomes us sum of the squares of standard normal deviate, but in this case it becomes a chi square. Obviously, as in the last case just the example we considered, but without any loss of degrees of freedom. So, it will be chi square with n degrees of freedom.

Now, when we want to find out the interval so obviously, the interval would be which I am again highlighting for the case of the standard deviation. So, it will be when we put we have not yet simplified it, it will be basically chi square n because it is n is a degree of freedom. Now here note down 1 thing till to be covered is 1 minus alpha suffix 1 and on to the right is n comma the suffixes are alpha comma alpha by 2, now the alpha 1 and alpha 2 have some relations with 2 alpha definitely, but we are assuming and rightly so the chi square is non symmetric, but we will make some assumptions further on which will again see it repeatedly and it says that it you understand the concepts. So, when we simplify the overall so I will remove the colour from here because, that was the initial discussion we wanted to do now the actual confidence interval now becomes this.

So, n s dash square by chi square n comma F alpha by 2 which is the lower value and the upper value is again n s dash squared chi square and 1 minus alpha by 2; now for simplicity to find out the shortest interval as if you read the lines which is given, we can basically assume alpha 1 is equal to alpha 2 is equal to alpha by 2 and we can solve the problems accordingly. So, hence the interval concepts with respect to the standard deviation on the variance the population becomes this n s square which remains same as this, 1 n s square again remain the same as the first one only note down the changes it will be chi square n degrees of freedom remains same there is no change, but alpha by 2 and 1 minus alpha by 2 gets replaced by the corresponding values which is alpha by 2 and 1 minus alpha by 2 because, we are assuming alpha 1 is equal to alpha 2 to find out the shortest distance as that is equal to alpha by 2.

(Refer Slide Time: 14:09)



So, this is done, so a done in the sense I have just disc will be coming back to that later on time and again now we want to find out the confidence interval with respect to the mean value the standard deviation, given the population mean is unknown. So, if the moment the population mean is unknown; obviously, we will know we would not be using the s dash, but we will using the s without the dash. So, in that case s dash becomes 1 1 by n minus 1 and the summation remains as it is, but only in the case here mu is being replaced by the sample mean.

So, once you do the same concept and go to the logic, so the ratios of the s square by sigma square and you can prove it becomes a chi square with n minus 1 degrees of freedom, once you replace it again remembering it will be again the same concept of alpha 1 and alpha 2 and so on and so forth. So, once you solve the problem it becomes this which is now note down the simple changes, in the initial case it was n here for the case when you want to find out something to do with the confidence interval of the of sigma square. So, it was n now it is n minus 1, in the previous case it was s dash square now it is s square by the way just store the information this small n which is the suffix for s basically denote the sample size nothing else, so do not be confused about that.

Now in the right hand side also I am only discussing the numerators, the right hand side also the same replacements has taken place logically n has been replaced by n minus 1, s dash square suffix n has been replaced by s without the dash square suffix n and in the denominator the things have changed accordingly, our firm by 2 remains as it is 1 minus alpha 2 by 2 remains as it, but the degrees of freedom which was n initially has changed to n minus 1. So, this would basically be the lower limit and this would be the upper limit and basically the sigma square value would always lie between them, with 1 minus alpha being the level of confidence.

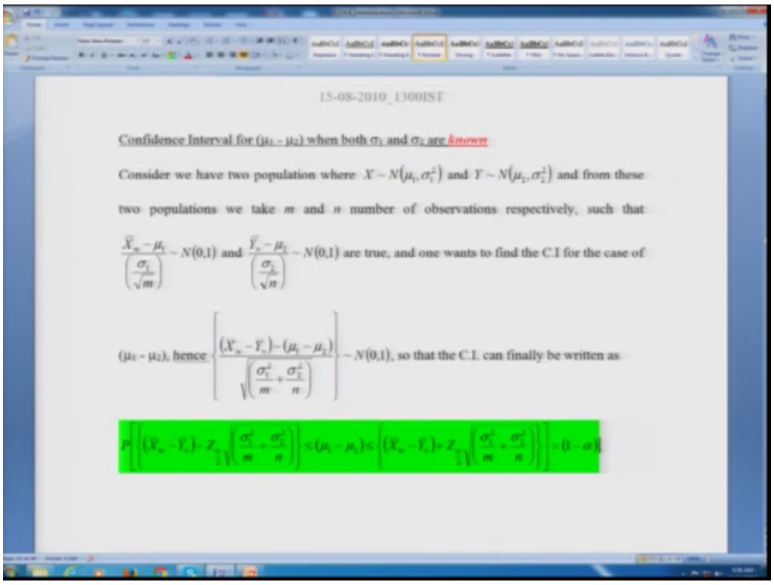
Now, I want to do something so let me pause here and let me discuss few examples I have not written them, but in order to make you understand. So, if you listen to me and you will understand, say for example in the first example where we had to do something with the mean value. So, you had basically 2 distributions z and t. So, those examples could be thought about as say for example, you are producing some goods in the factory and you want to find out what is the average weight of the jam bottles or the jelly bottles or some tie rods which are coming out what is the overall length; or say for example, in carton boxes are been packed what is the weights and all this thing or say for example, some very fine instrument is being produced. So, what is the precision of this instrument, what is the inner dimension of a tube which is coming out which is to be used for oil exploration. So, all these things can be thought about very simple examples.

So, when you want to do something to do with the average dimension we use basically the z or the t distribution, in the second case when we consider something to do with the variance it is to find out what is the variability or the dispersion of the average output or the value of the output which is happening; may be you are producing some goods again the same example packaging jam bottles jelly bottles packaging carton boxes trying to find out the (Refer Time: 17:46) or whatever we want to find out what is the variability. So, based on that we will use the chi square whatever the degrees of freedom would be it will be n or n minus 1 depending on the on information set which we have, information are I mean population mean being known population mean being not known.

Now, consider a slight change in the overall of the concept, you have 2 factories and the same products are being manufactured which will again repeat it is either filling up a jam bottle jelly bottle trying to find out the weights or trying to basically weigh the carton boxes which are being packed or trying to find out the length of the tie rod or trying to find out the inner dimension of a very fine pipe which will be utilized for oil drilling. Here the comparison is now being happening that 2 factories are producing 2 locations are producing, so what is the overall variability in the processes. So, if there is a variability. Obviously, we will consider the something to do with the variance. So, when you are trying to find out something to the various it will be something to do with the ratios of the variances, when it comes to the ratios of the variances obviously, you will see very logically that we will be using the F distribution.

Now if you remember the F distribution I did mention that, I remember that you basically try to use 2 degrees of freedom m and n, m and n being the sample size for the sample 1 and n being the sample size a sample 2. Now this degrees of freedom m n n will change, how will they change that is me proceed you to become creating you, so here is the background and now I am going to discuss.

(Refer Slide Time: 19:33)



So, the problem is I will highlight first, you want to find out something to do with the difference of the mean when both sigma 1 and square and sigma 2 are unknown are known just hang on. So, here what I what would be relevant for you is that first to discuss that concept or the differences of the mean rather than the variability's ratios, I will come to that availability issue because I thought that I have already discussed the example of when I when I did mention about trying to find out the weight of the jam bottles. So, rather than the variability let me put the same example in a different perspective the overall environment remains the same.

So, say for example, in the last example which I did miss did discuss you are trying to find out the company variability, rather than trying to find out the variability how about finding all the differences on the average values. So, this is the framework of the problem and let us see that how the concept of trying to find out something to do with the difference of the mean would be coming up. So, I will come to the F distribution later on please bear with me.

So, consider we have 2 populations X with certain mean and certain variance and the mean being mu suffix 1 when I mentioned 1 2 3 it will be with the suffix, if not I will basically specify it please and the variance is sigma square suffix n and y is again mu 2 sigma 2, 2 is the second 2 is basically for the suffix and from them you take a proper sample size of m as in mango n as in Nagpur, you take sample size of m n it can be n suffix 1 n suffix 2 it can be equal also it does not matter; number of observations have been taking up such that we know definitely because, if they are normal then obviously these 2 would definitely be true, one is the mean value of the first set of observation from the first population would follow a normal distribution with mu 1, as the mean value and sigma square suffix 1 divided by m as the variance.

So, based on that once you find out the distribution is this which you can understand. Now when we go to the second population take up a sample size of size n then why would base what is the average y bar suffix n, would be the average of the sample size it will be distributed with a normal distribution with a mean value of mu suffix 2 and variance of sigma square suffix 2 divided by n. So, this is what we have I will put a different colour here in order to make you understand.

So, these are the distributions for the sample means for the first sample means for the second; so if this is clear I will remove the colours and proceed further on. Now when you want to find out the difference of the sample mean it will be or to try to understand the difference for the population mean, what you will do is find out the difference between the sample mean compared with the population mean difference. Now any convex combination or normal distribution would always be normal.

So, in that and this is a very unique property of normal distribution. So, once you consider that the numerator of this now let me go very slowly, numerator of this would basically a difference of 2 terms the first term which is in the first bracket is the difference of the sample mean and the second term within this curve this curved bracket would basically be the population differences. So obviously, if the population differences are 0 then obviously, the sample mean differences should also be 0 technically.

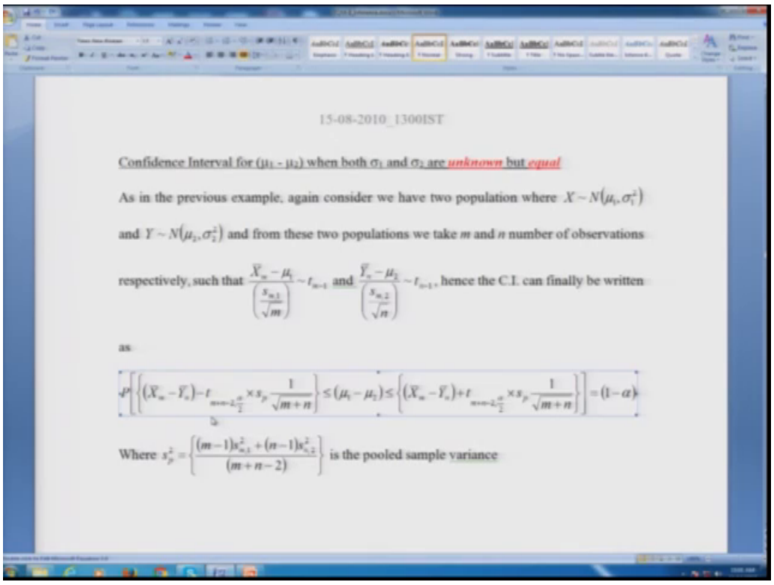
Now when I am trying to find out the differences this is the random variable, so obviously it means the differences or the 2 random variables would have a normal distribution and now the question is that what are their actual averages and variances; the average is mu suffix 1 minus mu suffix 2 and the actual variance of this difference between the sample mean would be sigma square suffix 1 divided by m plus sigma 2 surface square divided by n as shown here.

So, if I basically go into the actual concept of trying to find out the normal distribution which is now z, the pink coloured highlighted formula. So, this is basically the difference between the random variable in it is mean value divided by if you remember here this is divided by it is standard deviation. So, this is the standard division, so it will be easy normal 0 1. So, the confidence interval if you put in the formula it becomes I will just remove the colour in order to proceed and highlight it here, let me use some other colour which one the light green, so let us use the light green one.

So, is the probability of the average values of the difference of the average values for the 2 sample minus, now again see a very similar simple case in the other case when there was only 1 population it was basically z alpha suffix alpha by 2 into sigma which is the population mean divided by square root of n, because that was sigma square by n divided by n was the sample means variance.

So, in this case again the second part which is under the square root is basically the replacement of sigma by square root of n. So, that is less than equal to the sample population mean difference and again on the right hand side in place of minus, this what is the sign is there before z alpha by 2 it becomes plus z alpha by 2. So, these are the limits for the lower limit and the upper limit considering that you want to find out something to do with the difference in the population mean, using the sample means as the best estimate and obviously that would be equal to 1 minus alpha.

(Refer Slide Time: 25:50)



Now, when we want to find out something to do with the I will come to the F distribution please hang on, when you want to find out something to do with the difference in the population mean considering the standard deviation of the variances are unknown for the population but they are equal; then obviously, the question will come up that why did I mention that, now think intuitively and it will come out very easily. If the population if the variances are unknown so obviously, you to replace them with the corresponding values of the sample.

So, in the sample case so obviously this is true where it is being highlighted this is true which we have already discussed. Now very interestingly this would also be true because that is equal to the t distribution with m minus 1 and n minus 1 has a degrees of freedom considering you have lost 1 degrees of freedom in both the cases because, you are now using s without the dash in both the cases. So, s without dashes is basically the s m comma 1 and m comma 2 see this 1 and 2 basically signifies the from which cooperation you are picking up. So, these are not the s with the dash they are s without the dash.

So, once you plug in the sampled pooled sample variance would be given by the formula which I will try to highlight first, so which is the yellow colour. So, this is the pooled sample variance with the corresponding s das is coming for both the pop from the variances and the this total combined reduction in the degrees of freedom would be 2, 1 from the first population 1 from the second populations hence now total combined degrees of freedom is m plus n minus 2. So, once you have this so once you replace that in the in the limit confidence interval the formula become this, difference in the sample mean now in the other case when the population variance was not known you use the t distribution with 1 degrees of freedom being lost and then you proceeded, see here is exactly the same thing. This is t distribution obviously, with the minus n because this is the lower limit.

Now watch it very carefully in that case in the first case when there was only 1 population it is there was only 1 degrees of freedom loss which was m minus n or n minus 1, whatever was m minus 1 or n minus 1 here it is m plus n the combined sample size minus 2 because, you are losing one degrees of freedom in both. So, this is the suffix which are coming sp is the s suffix p this pooled variance and obviously, 1 by square root of m plus n and here on the right hand side also the same formula, but with a plus sign coming in front of t t and all the suffixes. So, once you solve it you are able to find out what is basically the confidence interval and solve your problems accordingly.

So, with this I will end the fifth lecture and continue discussing more about the confidence interval and hypothesis testing in the sixth lecture and so on and so forth. So, try to wrap up as soon as possible and try to continue the design of experiments. So, once you understand that it will be much easier for all of us to proceed a little bit fast, in the case of design of experiments and the concepts accordingly.

Thank you very much have a nice day.