Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 21
Inferential Analytics Part – 2

Hello everybody, this is Rudra Pradhan here and welcome you all to BMD lecture series. Today our discussion is inferential analytics and that two part two; in the last unit we have already highlighted the inferential analytics part one. And the idea behind the particular discussion was to solve some of the business problems relating to one sample case and the structure of the inferential analytics is to draw the kind of conclusion behind the population parameters.

And specifically, we like to address three things; first to know about the cost structure about the population parameters, that to check whether the structure of population parameter still valid or not valid by using some of the a sample informations.

And second thing is to know the confidence interval about the population parameter and then to find out the optimum sample size. And specifically, we have solved all these problems and discussed all these three issues relating to the comment on population parameter and the confidence interval about the population parameters and the optimum size of samples and that too by using a; Z statistic, t statistics and chi square statistic.

And in this particular unit, we will address all these problems and by using the kind of multivariate structures. So that means, in this particular inferential analytics that is on part two; we will solve some of the business problems relating to multivariate case.

So; that means, at least we must have two different samples or two different variables through which you have to draw the conclusions and that too comment on the population parameters and again the confidence interval about the population parameter and the optimum sample size. So in fact, in this case we will address one more thing that is on F statistic; which we have not covered in inferential analytics because F statistic is typically multivariate kind of; means it will be applied in a kind of in a multivariate situation.

So, corresponding to this particular brief structure.

(Refer Slide Time: 02:41)
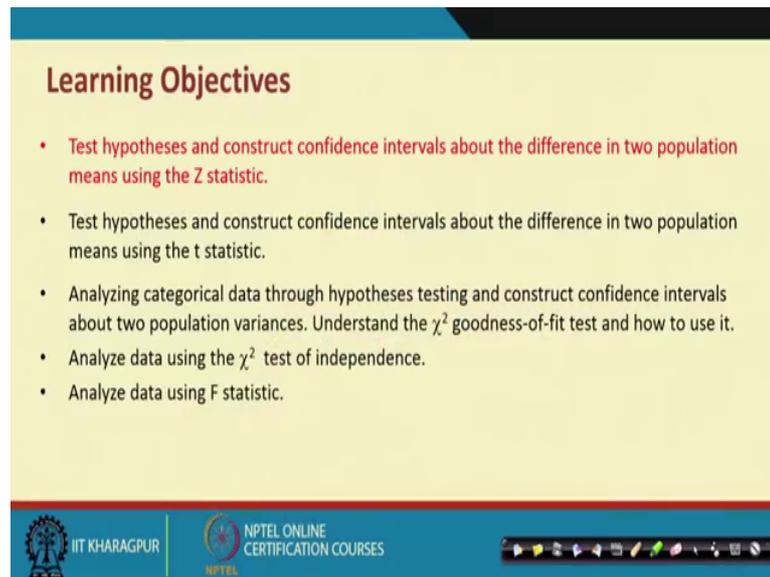
So, our typical objective is here to test hypothesis and construct confidence intervals about the difference in two population means and using these Z statistics. Again to test the hypothesis and construct confidence interval about the difference in two population means by using the t statistics. And again we solve some of the problems relating to categorical data and by using chi square statistics and then we resolve some of the problems by using F statistics.

So, typically we are going to address t statistics, Z statistics, chi square statistics and F statistics. And then we will apply statistic in the situation where at least we have two sample case or multiple sample case. So, initially we start with two sample case; then we extend to multiple sample case. So, before we start the process; I like to highlight that any kind of business analytics, investigation or statistical investigations.

So, the structure of testing with respect to one samples, two samples and multiple samples; that means, univariate structure, bivariate structure and multivariate structure. In the univariate structure, we must have one variables and that to one sample through which we have to draw the conclusion about the population parameter and bring the confidence interval about the population parameter. And then we have to find out optimum sample size corresponding to a particular test statistics.

And in the bivariate case, you must have a two samples and then we like to know the difference about this; maybe mean difference or variance difference corresponding to

two samples. And then we like to draw the conclusions about the population parameter, then again confidence interval about the population parameter and the size of these samples.

And then finally, we will go for multivariate case where we must have multiple sample case; that means, it is a multiple very variable case through which you have to address the business problems. So, if you go to the kind of in a business environment; some of the problems can be connected with the univariate structure, some of the problems can be connected with the bivariate structure and some of the problems can be connected with in multivariate structures.

So, the univariate structure which we have already discussed in the case of inferential analytics 1. And here we are going to discuss these same problems with in multiple case; that is in we start with the bivariate case first then we will connect with multivariate case.

(Refer Slide Time: 05:36)

So, the highlights of the lecture is like this; we will again go by a sample versus populations, then multivariate problems and population proportions, population variance then again we will use all these test statistic, Z statistic, t statistics, chi square statistics and F statistics and then we will connect with a case called as ANOVA and categorical data analysis and finally, we have to go for goodness fit structure.

(Refer Slide Time: 06:06)
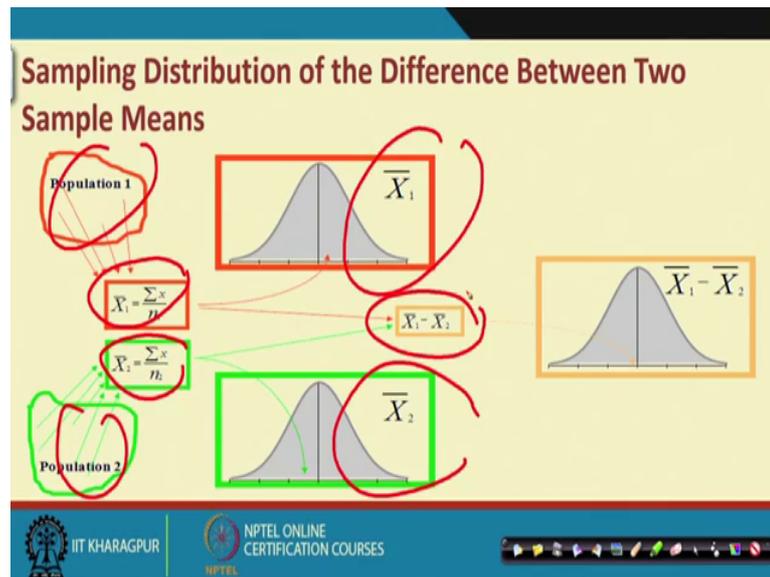
And in the mean times, in this particular lecture typically I will address the same problem through Z statistics; that means, the corresponding to inferential analytics once, we have already discussed some of the problems connecting to Z statistics that to comment on the population parameters. Then to know the confidence interval about the population parameters and then optimum sample size of the particular test statistics.

But here it is the same case, but we are dealing with at least two samples so; that means, the structure of the problem is you must have a two populations. And you are supposed to draw two samples from the two population and with the assumption that the both the population are more or less same. And then the sample difference will find out and check with population difference that is the population parameters kind of difference. So, here is the typical structure is like this.

So, there is a population 1 here typically; so this is population 1 and this is a population 2. And in the population 1's we are drawing a particular samples and the corresponding sample mean is X bar. And in the population 2; we are drawing a particular sample called as a X 2 bar; X 2 and corresponding to X 2 samples. So, our sample mean will be X 2 bars. Now; so, we are drawing a sample X 1 from the population 1 and against we are drawing a sample X 2 from population 2. And with the assumption that the population parameter for first one is mu 1 and population parameter for second one; is nothing, but mu 2.

So, now the idea behind this particular structure is we like to check; what is the difference between these two population parameter? And then we will draw the sample statistic and then mean of the sample statistics and we try to find out their difference. Now the idea behind this particular structure is to check whether this mean of the sample difference will be converged to the mean of these two populations.

So, with this objective; we will be proceed to this particular discussions. So, now corresponding to the discussion process; so, this is how the first case and this is how the second case. And by the way, the difference between these two; sample mean is nothing but called as X 1 bar minus X 2 bar.

So, now earlier discussions we liked to check the mean of the particular samples and then mean of the population; then we are checking the difference. So, here we are checking the mean of the two samples and then we are comparing with the difference of the two populations. So, then we would like to know how is this particular structure all together? So with this; so, let me show you how it can be.

(Refer Slide Time: 09:12)
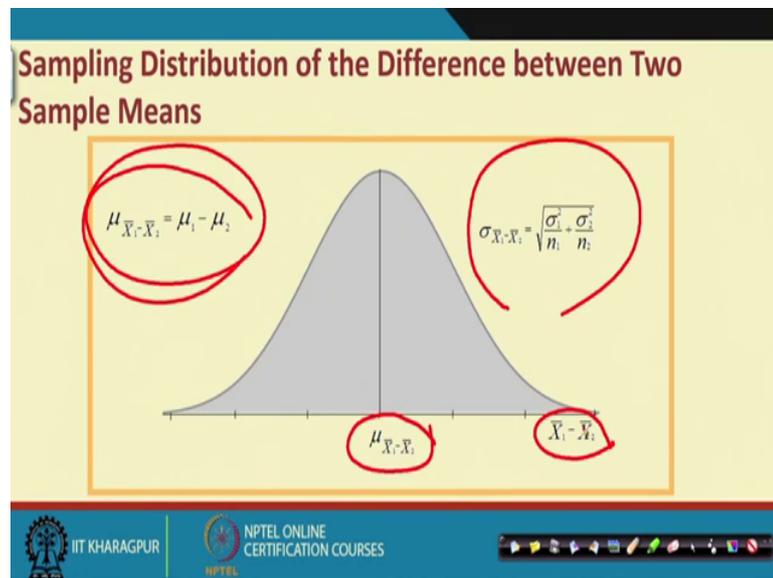
So this is how the particular structure and so corresponding to this. So, this is how the particular structure; so, here the idea being this particular structure is; it is a population mean difference that is nothing, but college mu 1 minus mu 2 and mu 1 is a population parameter for population 1 and mu 2 is the mean of the population 2.

So, this is a population parameter for population 1; so, here the standardization is nothing, but actually for two samples will be sigma 1 square by n 1 and sigma 2 square by n 2; and that too square root of both. And we like to check actually, this is the population mean and this is what the sample means. We like to check; what is the difference between these two? Means population mean difference and then sample mean difference.

And our idea is to check whether both are same or there is a kind of difference. If there is a kind of difference, then we like to check what is the statistical kind of difference? So, if statistically the difference is significant; then this problem can be addressed differently. If the difference is not statistically different, then we will not bother about the particular issue.

But here our job is to check whether there is a difference and if there is a difference whether the difference is statistically significant or not. And as a results, we will be dealing with hypothesis testing; it is the same structures we have to fix the probability levels; then we have to create a confidence intervals. And then we have to see whether the calculated test statistics will be over take the kind of tabulated statistics.

And if that is the case, then we are going to reject the true null hypothesis; otherwise we will accept the true null hypothesis. By the way, here the null hypothesis that there is no difference between these two population means and the alternative hypothesis will be there is a difference. So, accordingly; so the corresponding structure will be like this.

(Refer Slide Time: 11:20)

So, this is how the; first we start with the particular structure by Z statistic; then again we need do by t statistics. So, now as per the earlier discussion in inferential analytics 1 and we can apply Z in two different situations. In one situation, your sample size should be high that is greater than 230 and the samples will be a mean in the case of particular samples; your mean is a population parameter sigma is given.

And sometimes population parameter is not given so; that means, typically in the case of Z statistic which you have already discussed in inferential analytics 1 and that was at least, you your sample size should be greater than 230. And then we have a two different option in one of since your population standard deviation is given; in another situation your population standard deviation is not given.

So, if it is given then evenly use and according the test it, if not given then we will use sample information as a proxy to know the variance and then we go for unit testing. And in this case we are dealing with the same structure, but here there are two samples and in the both the cases; we are assuming that sample size should be greater than 230.

So, that is how; so, the structure is here n 1 greater than 230 and n 2 greater than 230; that is the condition and then another assumption we like to put here before this plus testing procedure; so, they are independent. So, they are two in independent samples and then accordingly we would like to check whether there is a kind of difference between the two. So, accordingly so, the Z distribution will be like this means it is a Z statistic.

Earlier, we are discussing the Z statistic for one sample case is nothing, but X bar minus mu divided by sigma by root n. And here it is the same structure, but we are restructuring with respect to two samples and in this case instead of X bar. So, it is the difference between X 1 bar minus X 2 and here in the case of mu; we are putting in mu 1 minus mu 2 that is the difference between two population parameters.

And then instead of sigma by a root n; so, we have a sigma 1 square that is for population 1 and sigma 2 square for population 2. And the sample one is the first population and the sample n 2 is the sample observation for the second populations. So, as a result; so the Z will be it s a same structure, but it is a little bit restructures as per the particular requirement and that too with the two sample case, where we have a two different sample means and two different population parameters and population mean and then we have a population standard deviation.

Accordingly, we are just restructuring the particular process so, then accordingly will move to this particular process.

(Refer Slide Time: 14:23)
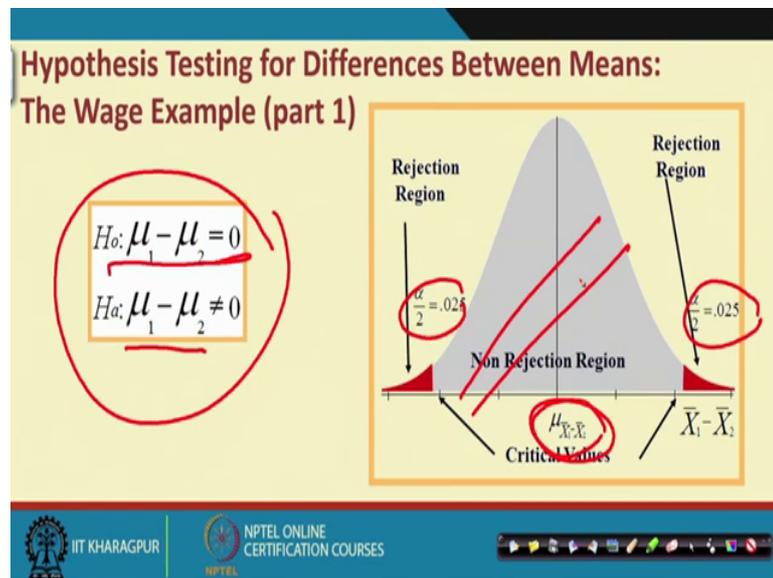
And as we have already highlighted that in this particular testing procedure, your null hypothesis will be mu 1 equal to mu 2. And as a result the difference between the two equal to 0 so; that means, mu 1 minus mu 2 equal to 0 and alternative hypothesis will be mu 1 not equal to mu 2 and as a result the difference between these two will be not equal to 0.

So, then the test procedure is almost all same; so, this is what they actually a population mean difference and then this is the sample difference. And we like to check whether the sample difference is different from the population mean difference. So, as a result the first step is to test the means fix the alpha value and you have to target one tail or two tail. Then in this case we are putting actually two tail so that fixing 5 percent then.025 will come to this side and.025 will come to this side.

And this is what actually the acceptance zones or non rejections zones. So, now, we like to check where is the position of power calculated value? So, let us take an example and then we will address this problem.

(Refer Slide Time: 15:39)
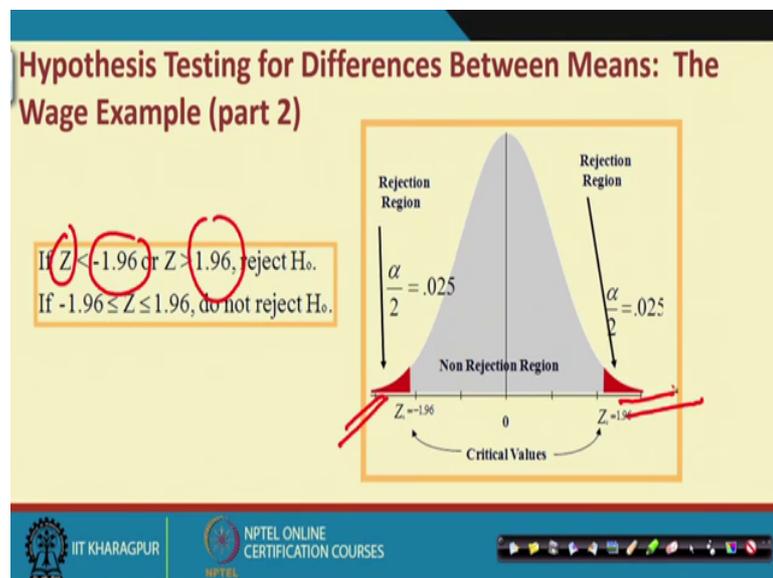
So, in the examples; so, we are fixing actually as per the kind of alpha 5 percent and that to for two tail. So, your Z critical will be 1.96 and a left side the value will be minus 1.96 and for the right side we will put actually 1.96. So, now we have to reject h 0 means null hypothesis if the Z calculated will overtake the Z critical so; that means, it should be fall under this zones or it should be fall under this.

So, accordingly we have to take a decisions and this is how the kind of in confidence interval structure. And after placing the test statistics, we will accomplish conclusions whether there is a difference of sample means and if there is a difference of sample means and how this difference is actually convergence towards the population mean

difference? And if there is a difference and whether it is statistically different, so these are the things we are supposed to check hear and then accordingly we are addressing.

(Refer Slide Time: 16:46)

So, now for the sake of understanding; so, we are taking two different actually data set. And this is the case of sample 1 and this is the case of sample 2 and as a result; in the first case sample size is 32; that is n 1 and curve using all the sample points, we will calculate the sample mean; that is X 1 bars.

And in this case, it is coming 70.7 and the corresponding sample variance is a 16.253 and this is sample variance and this is standard deviations. And accordingly in the case of second sample, so the sample size is 34.

So, mean is 62.187, then sample variance is 166.411 and the corresponding standard deviation will be 12.9 and accordingly we will actually test the particular process. So, let us go to these particular structures and then we will see the kind of result.
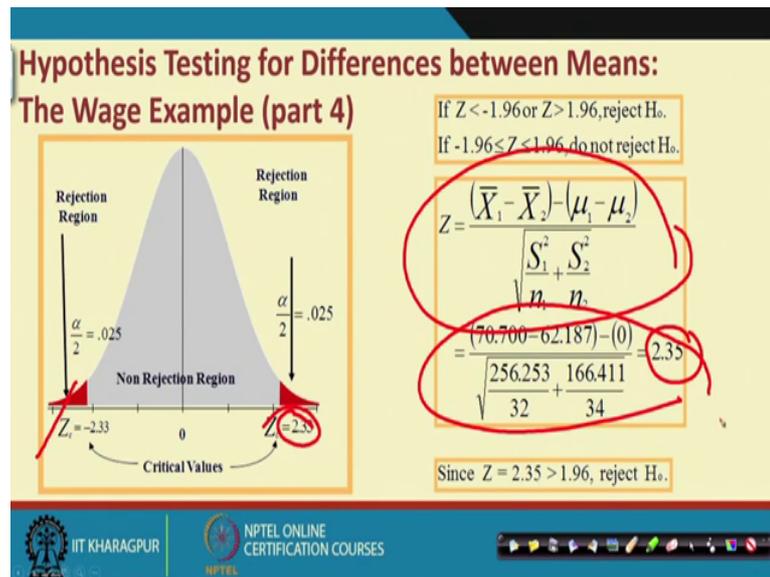
So, now, using the sample informations then X 1 bar is already reported, X 2 bar already reported, S 1 square is already reported, S 2 square is already reported, so; that means, previous slides we have already reported everything, then accordingly we are just putting all these value to these Z statistics. And then we like to check the kind of results; this is the Z statistics formula and then we are putting these values only.

And after that we are getting the Z structure; Z value is nothing, but called as 2.35 after simplification. And then as this is what actually called as a Z calculated; now what we will it do actually. So, we like to take this Z calculated to these kind of in a confidence interval and you check the position of calculated Z, whether it is coming under the rejected zones or whether it is coming under accepted zone.

Since it is coming actually a 2.35 and as a result, so it will be coming in the right side and here the Z critical actually showing 2.33 and this value is coming 2.35. So; that means, it is actually more than that; so as a results. So, we are in a position to reject this true null hypothesis. So, this is how the kind of structures we have to reject the true null hypothesis.

So, now corresponding to this is; so, means corresponding to this problem. So, we are here to draw the conclusion that; so corresponding to this problem. So, we are actually finding the situation that there is a difference of the sample mean and the sample difference is actually a not converging towards population difference. And the sample

mean difference and population mean difference, there is a gap and that gap is a found to be statistically significantly here.
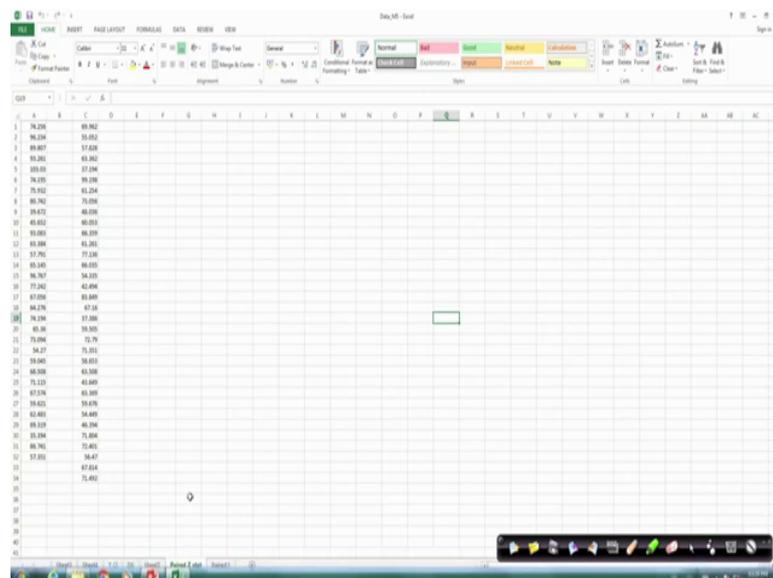
(Refer Slide Time: 19:50)

And accordingly so, the business decision need to be taken care of. So, this is actually a through excel spreadsheet and that we have to draw to get the kind of conclusions. So, let me take you to this kind of problems, then you highlight how this problem can be solved through actually the kind of structure.

(Refer Slide Time: 20:09)

This is what actually the kind of structures and this is this data set and with this data set; so, what will you do? So, we will just actually calculated the kind of difference and then with the help of this difference. So, we have to check whether it is statistically showing some kind of significant or not. The same data which you have here; so, we are actually plotting and then we are calculating the kind of mean structures.

So, what will it do actually? So, you go to this particular data analysis and then; so, this is what actually. So, let me take you to this particular structure; so, here actually the kind of difference and whatever samples are here. So, what will you do? So, corresponding to this samples; so, you just report here and if you go by manually then you put here find out this average.

So, it will be appearing average here and against you put here average of second one; you find out the average of second ones. So, this is coming actually 62 percent then corresponding to this problem; if you will find here. So, it is actually 70.7 here and then it is actually coming 62.12.

So, then against you go to this data and instead of mean; so, you find out the variance. So, against you go to this particular structure and try to find out standard deviation of the series. So, it is coming 16.25 and against you find out the standard deviation of the second series and against you find out here. So, this is what actual standard deviation of second series; then again if you go to this samples summary, so it is actually matching.

So; that means from this series you have to calculate the mean of the series and variants of the series and then sample size; that is n equal to 32. And again in this case, so you have to find out the mean of this particular series, variance of this particular series and then the kind of sample size.

So, now after getting all these sample size and mean and variance; so you have to actually just put in the Z. And then the Z value will give you the kind of structure; the exact value. So, here there are two ways you can compare the particular results; in one case what is it do? So, since Z is coming 2.35.

So, now go to the Z tables and find out what is the probability value for Z equal to 2.35 so; that means through p value approach; you can actually take the decisions. If not then fixing alpha and for instance here you are fixing alpha; you are getting actually means

0.5 percent; you are getting the Z critical is 1.96. So, either you compare with Z critical; with the z calculates or you can fix alpha equal to some things; in this case 5 percent, then you find out here at Z equal to 2.35; what is the probability structure?

So, then you will compare whether it is coming under the rejected zones or coming under the accepted zones. So, as a results accordingly you have to find out the kind of you know difference. And then the particular difference, we can check whether it is a; means we have to take the decision whether the particular difference is coming statistically significant or not.

So, now in the excel spreadsheet whatever information we have already gathered so just you to summarize here. And this is actually the mean of this first series, mean of the second series, variance of the first series, variance of the second series and n 1 is 32 and pg 34. And the mean difference we are assuming that here mu 1 equal to mu 2 and that is population 1 and population 2. So, as a result the difference would be equal to 0 and Z we have already calculated that is 2.35. Now corresponding to Z 2.35; so you are getting actually probability level at one tells 0.009.

So now if you put in alpha equal to 5 percent and by default; so, it is coming under actually rejected journal. So, as a results; so, we will come to the conclusion that there is a significant difference between these two population series and the main difference of the samples are not converging the population mean difference. So, as a result; so, the problem need to be investigated further to get some kind of inference for the kind of as per the particular requirement.

(Refer Slide Time: 25:22)
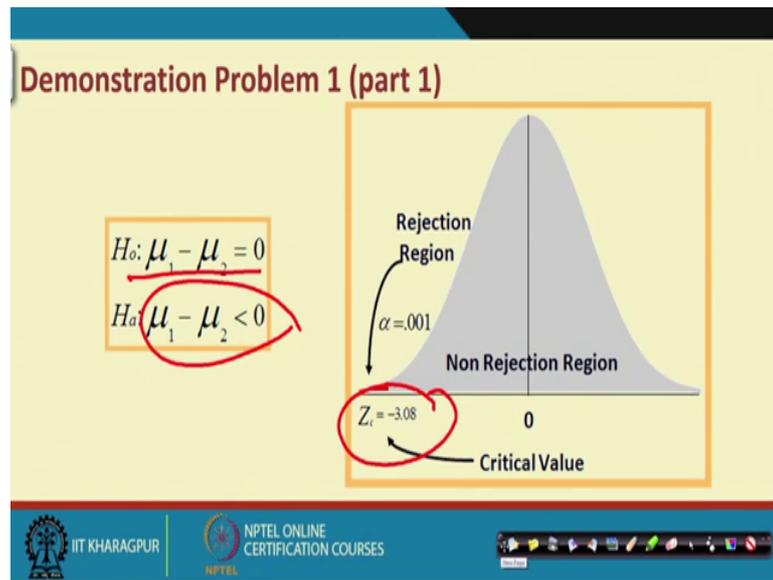
Now against there are two typical case, if in the first case mu 1 minus mu 2 equal to 0; that is the null hypothesis. And then we are fixing alternative mu 1 less than equal to mu 2. So, as a result; so like this, so we are putting first mu 1 equal to mu 2. So, as a result the difference between the two equal to 0 and in other case if not alternative by default will be mu 1 not equal to mu 2.

So, now if mu are not equal to mu 2 then will go by both the sides. So, now we can go for one tail test also; in that case either we will fix actually mu 1 less than equal to mu 2 or we will be fixed mu 2 less than equal to mu 1. So, now, if mu 1 less than 2 mu 2; then mu 1 minus mu 2 is less than to 0; so, as a result, the corresponding to actually alpha 1 percent; then Z critical will be coming actually minus 3.08. Then again, we have to calculate the Z value and then we will see whether the Z calculated is overtaking the Z criticals or under the; means that is actually a critical value, whether it is crossing the Z tabulated value or it is under Z tabulated, then accordingly will it take a decisions. So, this is what actually the; so, this is the case one where you know mu 1 less than to mu 2 and then.

(Refer Slide Time: 26:55)

So, accordingly; so, the problem is here, so in the first case when mu 1 less than to mu 2. So, the same structures, but it is with a different problem and again there are two samples and X one bar is the first sample mean. And then this is a standard deviation is a S 1, then corresponding to sample 1, then sample 2 mean is equal to this much X 2 bar; then it is the standard deviation S 2 and sample size is a 87 and sample size is 76.

So, accordingly you put all these value and then report the Z statistics and now the Z critical is coming actually minus 10.42. And corresponding alpha equal to 1 percent and one tail; then the Z critical is coming actually minus 3.08, but here your Z value is coming minus 10.42. So, it is not actually crossing the Z critical; so, as a result; so, we are in a position to reject the H 0. So, as a results the conclusion to reject the H 0 so; that means, there is a difference and the difference is statistically significant.

(Refer Slide Time: 28:13)

So, now in these particular structures having the kind of Z structure, so we like to highlight that there can be actually to calculate the confidence intervals. Earlier, we have discussed the concept of confidence intervals; so, that is the case of one sample case. Now, we have the case of two sample case that to the difference so; that means, actually; so, in the one sample case the structure of confidence interval will be X bar plus, minus Z into sigma by root n and X plus Z into sigma by root n; will give you upper bound or a upper critical value. X bar minus Z cross sigma by root n will give you the lower bound.

So, now in the same way; so, here instead of X bar; so, we are reporting actually X 1 bar minus X 2 bar that is the difference between two sample mean. And then; so in place of Z sigma root n, so we are putting the standard error of the mean difference; that is coming Z square root of S 1 square by n 1 plus S 2 square by n 2. So, again; so, taking the difference of the sample means and the Z value, then accordingly you will find the lower bound and find the upper bound.

So, typically this representation will be like this. So, this is actually probability of all these confidence interval with nothing, but actually called as 1 minus alpha. Because alpha by 2; we are putting actually in both the sides and then we are creating a kind of confidence interval by the total kind of probability that is 1 minus alpha. So, in the next slides I will show you the kind of structure.

(Refer Slide Time: 30:26)

So, this is what the kind of structure and in this case again we are taking two different samples. And the sample size; this is actually samples sample information n 1 is 50 and 2 equal to 50. And X 1 bar equal to coming to 2145 and X 2 bar is a 24.6, then the standard deviation and the standard deviation.

And then 95 percent confidence interval, it is coming actually 1.96. So, now what is actually the standard structure? So, you have to just find out the mean difference between the two samples and then we have to just calculate the standard error; that is Z upon square root of S 1 square by n 1 plus S 2 square by n 2; then after putting all these value. So, you are in a position to report that the confidence interval for mu 1 minus mu 2 will be minus 4.42 and minus 1.88.

Corresponding into the situation that confidence interval will be 95 percent so; that means, actually in this case mu 1 minus mu 2; that is the population mean difference which is actually ranging between minus 4.42 to minus 1.88. So, this is actually similar to the case which you have discussed in inferential analytics 1. And here we have actually two different sample case that is why the confidence interval will be created by taking the difference between the two samples only. And depending upon the fixing of confidence structures, we can create a confidence interval so; that means, in the particular structure what is the kind of one minus alpha.
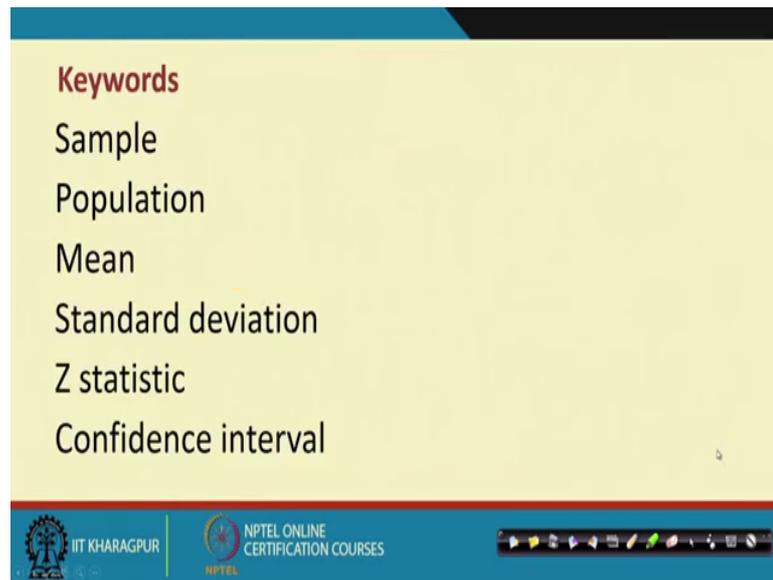
So, now what we have actually discussed in this particular inferential analytics 2. So, we are actually addressing the same issue; so that is the comment about the population parameter by using the sample information. And then we are targeting to find out the confidence interval for the population parameter and again targeting to find out the optimum sample size.

And here the case is actually two sample case; so that means, the assumption is that there is there are two different samples and they are drawing from the two different population. So, as a result there is a information about two different population information that is the population parameters mu 1 and mu 2 respectively and their corresponding standard deviation will be also given.

Then with the help of population parameters; so, you have to draw two samples and then you have to find our sample statistics and the job is to check whether the mean of these two sample statistics are converging towards the mean of these two population parameters.

And accordingly, we have discussed the problem through Z statistics and at the same times we have already also discussed the concept like to create a confidence interval of this population parameter; that is the difference between the two population mean. So, by using the sample statistics then you can create the confidence interval; interval for these population parameter.

And in the next class, we will discuss the same structures by using the t distribution or t statistic and chi square statistic; with this we will stop here.

Thank you very much, Have a nice time.