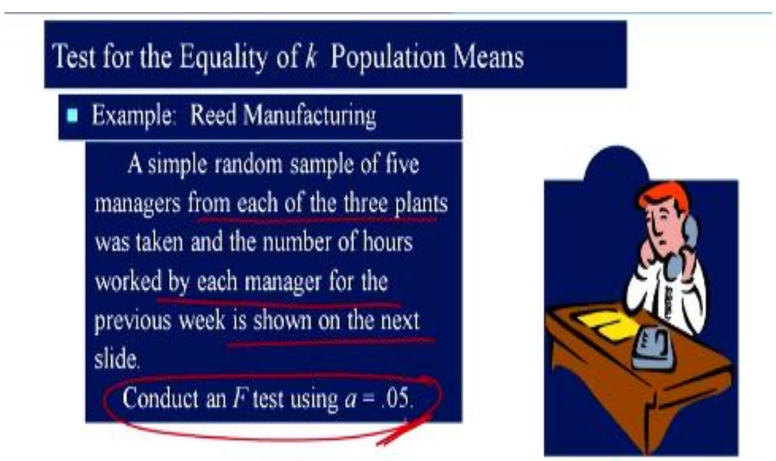
**Business Statistics**
**Prof. M. K. Barua**
**Department of Management Studies**
**Indian Institute of Technology - Roorkee**

**Lecture – 48**
**Analysis of Variance - III**

Hello friends, I welcome you all in this session, as you are aware in previous session we were discussing about ANOVA in we did work out couple of examples, in today's class, we will see some more examples ANOVA, so far we have seen in ANOVA whether there was significant difference exist or not and in most of the examples, we have solved we found that, we rejected null hypothesis in all those examples.

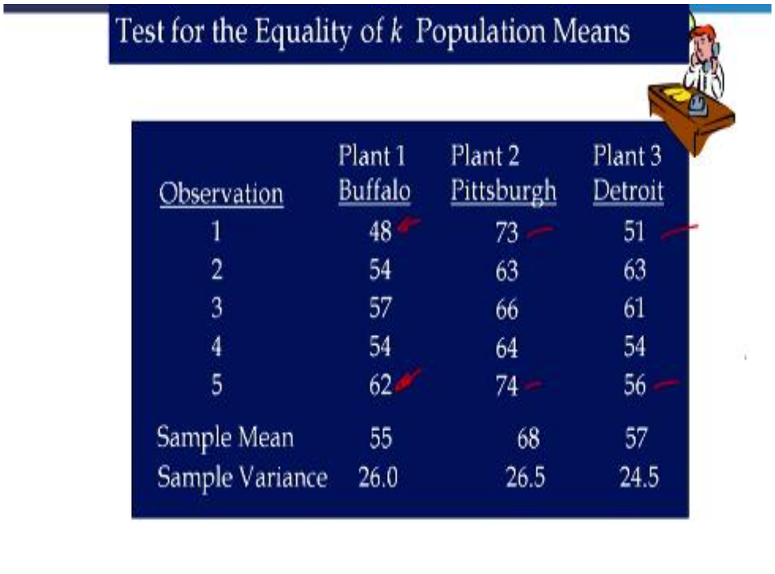
And we said that there was significant difference amongst different groups but we did not find amongst how many groups or between which 2 groups there is significant difference, so today we will see 2 methods, they are called Tukey test and Tukey Kramer test but before going on those 2 tests, let us look at more example on ANOVA, quite a simple example.

**(Refer Slide Time: 01:39)**




So, a simple random sample of 5 managers from each of three plants was taken and the number of hours worked by each managers for the previous week is shown in the next slide and we have to conduct an F test analysis are ANOVA at significance level of 0.05.

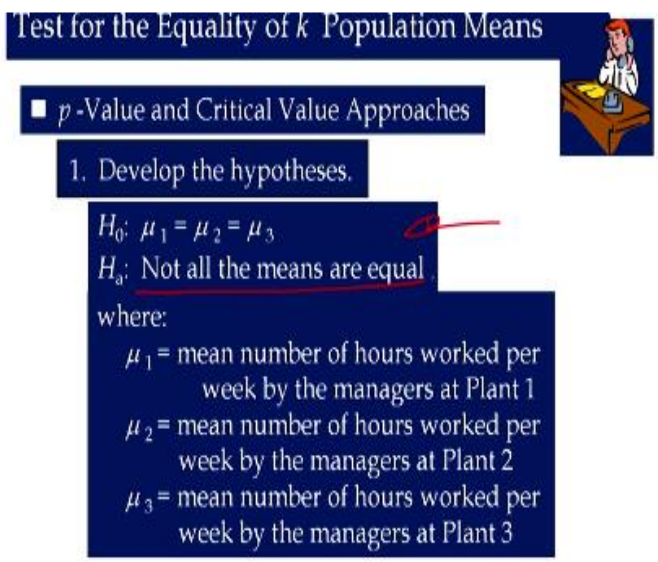
**(Refer Slide Time: 02:04)**




So, you have got this information, there are three plants; buffalo plant, Pittsburgh and Detroit, so these are the number of hours managers worked, okay, so first observation 48 hours, fifth observation 62 hours for plant buffalo for plant 2 73 and 74, for plant 3, Detroit 51, 56, so you need to find out is there are any significant difference amongst working hours, so far as these plants are concerned, right.

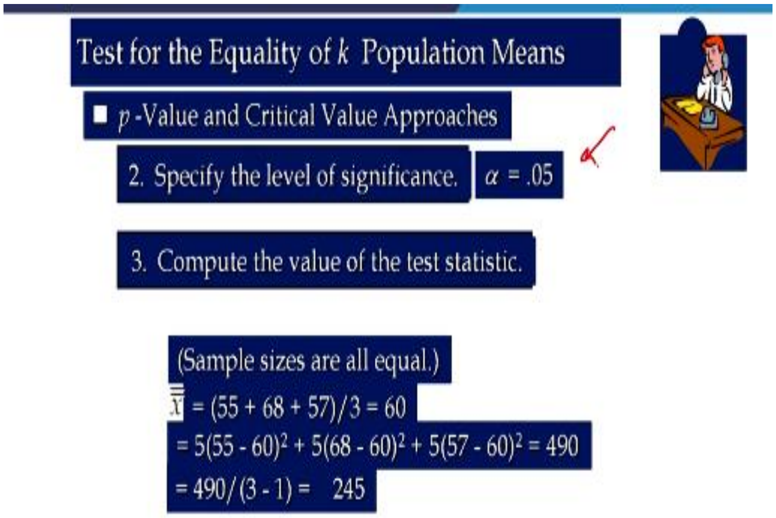
So, we will find out sample mean and of course, sum of the squared error and sum of squared among variance, right or in other words among column variance and within column variance.

**(Refer Slide Time: 03:06)**



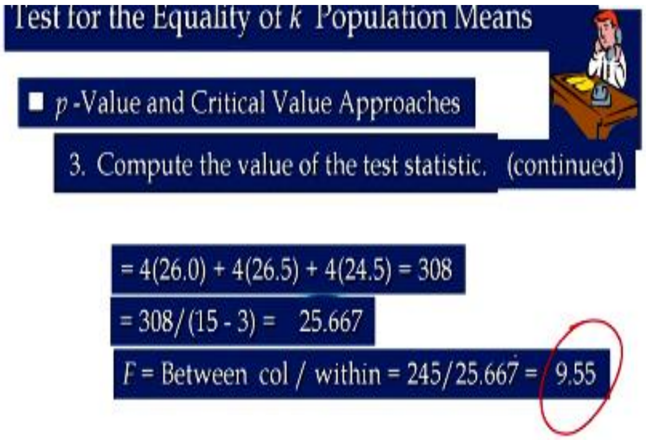
So, this how we can calculate it and this would be our null hypothesis that the mean number of hours worked per week by the managers of all these three plants are same and not all of them are same is alternative hypothesis.

**(Refer Slide Time: 03:25)**



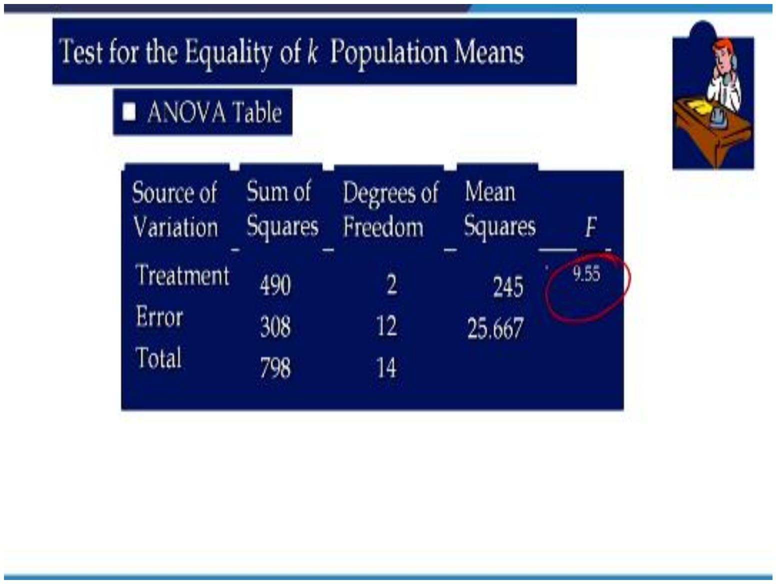

We know that our significance level is this, so we will calculate all means and grand mean.

**(Refer Slide Time: 03:36)**



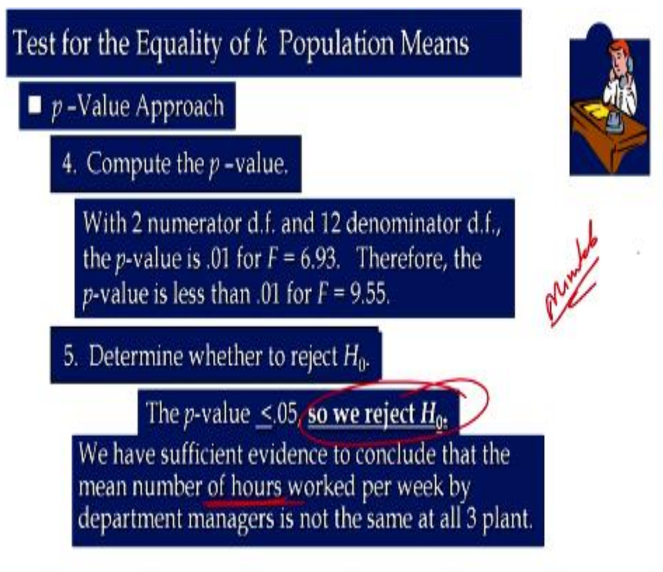

And then we will find out the F value which is 9.55, right.

**(Refer Slide Time: 03:46)**



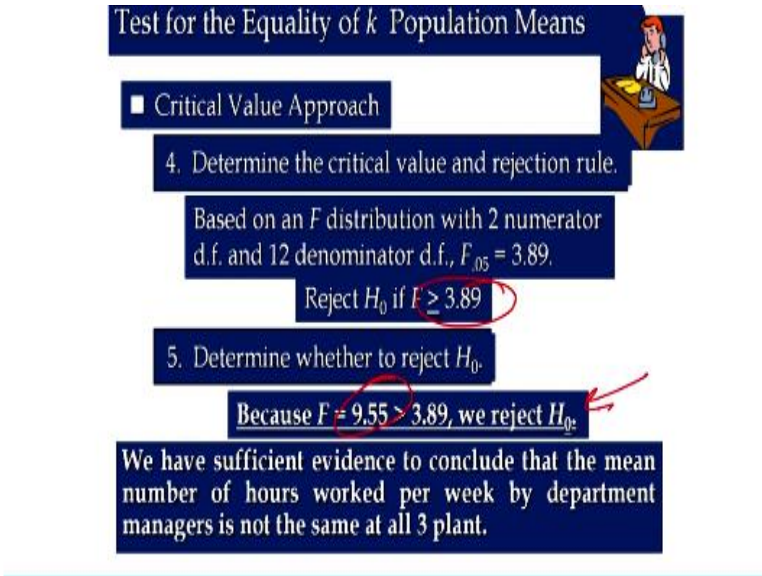

And then we will look at the table value and will compare the calculated F value with the table values, so here F value is 9.55, right.

**(Refer Slide Time: 03:57)**



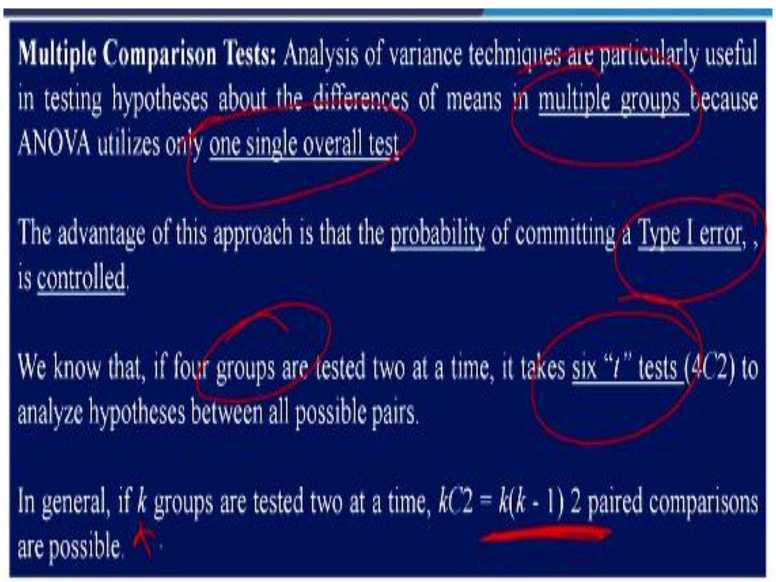

So, if you look at this question, then we will reject the null hypothesis and we will say that there is a significant difference in working hours, right number of hours worked per week, so this is a simple question which you can worked out using minitab software or any other software.

**(Refer Slide Time: 04:32)**




Let us move on to next slide, in fact you can solve this question using critical value approach as well, so this was our table value and the calculated value was 9.55, so we will reject null hypothesis, right.

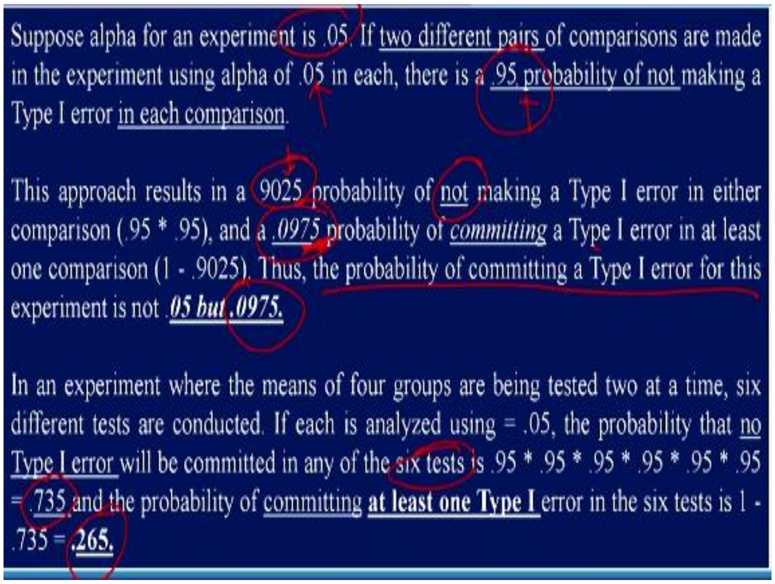
**(Refer Slide Time: 04:48)**




Now, let us look at multiple comparison test, as I now mention that ANOVA tells us whether significance difference exists amongst groups or not but it does not tell significant difference exist between which 2 groups or amongst which 3 are more groups, right, so let is look at multiple comparison tests, we know that ANOVA is the technique, it is particularly useful in testing hypothesis of differences of means in multiple groups.

Because ANOVA utilises one single overall text rather than having different paired t test or paired Z test, right, so the advantage of this approach is that the probability of committing type I error is minimised or it is controlled, so if there are 4 groups, we will have to have 6 different pairs of t test, right, so this how you can have number of pairs, if k is the number of groups.

**(Refer Slide Time: 06:05)**

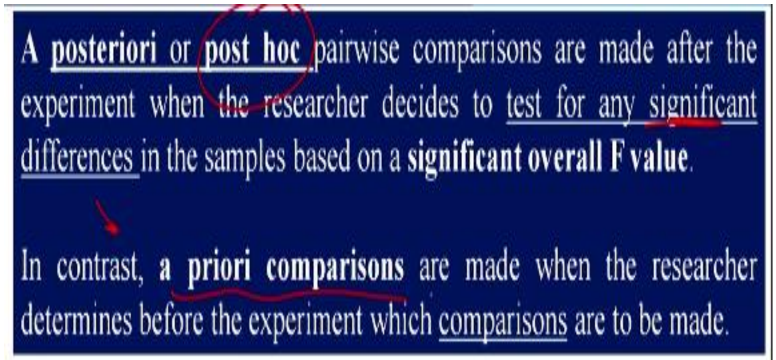



In fact this is what we have already discussed but let me tell you once again, we know that let us say if alpha is 0.05, so if 2 different pairs of comparison are made in experiment using this value, then there is 0.95 probability of not rejecting, not committing type I error, right, in other words let us say, the probability of not making type I error is this and let me tell you once again, so this is the probability of not making type I error.

So, if there are 2 pairs, so this is the probability of not making type I error, right, this is probability of not making type I error, it means this is the probability of making type I error, so 1 - this is this, right, this is the probability of making type I error, so we can say that the probability of committing type I error for this experiment is not 0.05 but it is only 0.097, in other words it is approximately 10%, right, okay.

So, if we have 4 groups, then you would be doing 6 t tests and the probability of not making type I error is this and the probability of making type I error is this which is very high, right, so that is why ANOVA is an appropriate technique because it controls what; it controls type 1 error, right.

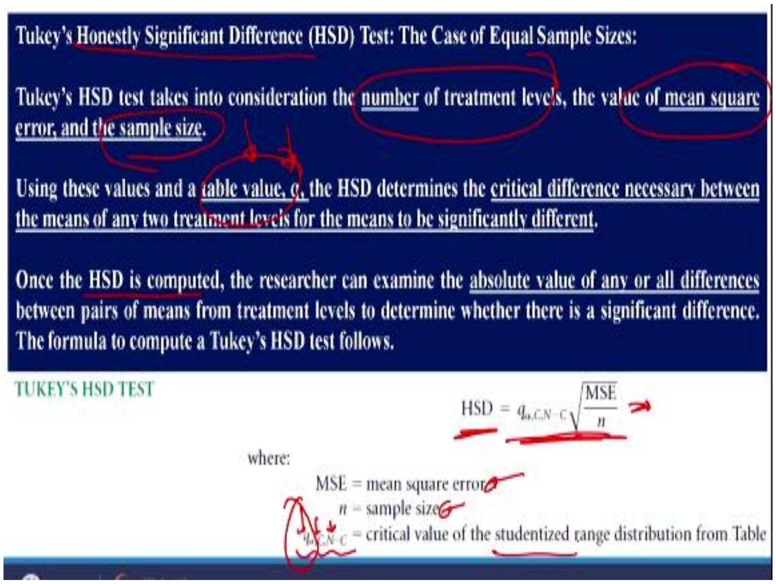
**(Refer Slide Time: 08:20)**




Let us look at and there is something called posteriori and post hoc comparison, so post hoc comparisons are made after the experiment when researchers decide to test for any significant differences, so as I said ANOVA does not tell you within which 2 groups there is a significant difference but there is something called post hoc test, it will tell you within or between which 2 groups, there is significant difference.

Apart from post hoc test, you have heard just something called a priori comparison which will help us in deciding the comparison is to be made between which 2 groups or which comparisons are to be met.

**(Refer Slide Time: 09:16)**

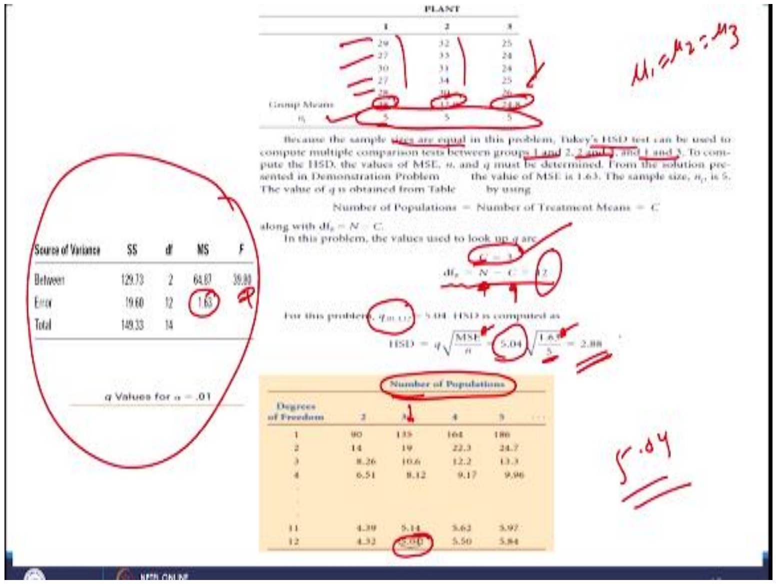



So, let us look at first test, it is called Tukey's HSD test or Tukey's honestly significant difference test, now this test is applicable to know significant difference amongst groups when the sample size of each group is same, it takes into account 3 things, it takes into account the number of treatment level, the value of mean square error and sample size, right. So, if we all these 3, we will calculate HSD value.

And for that we would be requiring something called table value q, so this HSD determines the be critical value from table and we compare that critical value with this calculated HSD value, so the HSD determines the critical difference necessary between the means of 2 treatment levels for the means to be significantly different, so once this is computed, you are comparing this computed value with the table value at an appropriate degrees of freedom.

So, this is nothing, it is under root of MSE/n, so this mean square error, n is sample size and this is what is q alpha, c and n – c, so we have to see what is the value of q alpha at c and n - c degrees of freedom in studentized table, so this is known as Tukey's HSD; honestly significant difference test.

**(Refer Slide Time: 11:14)**



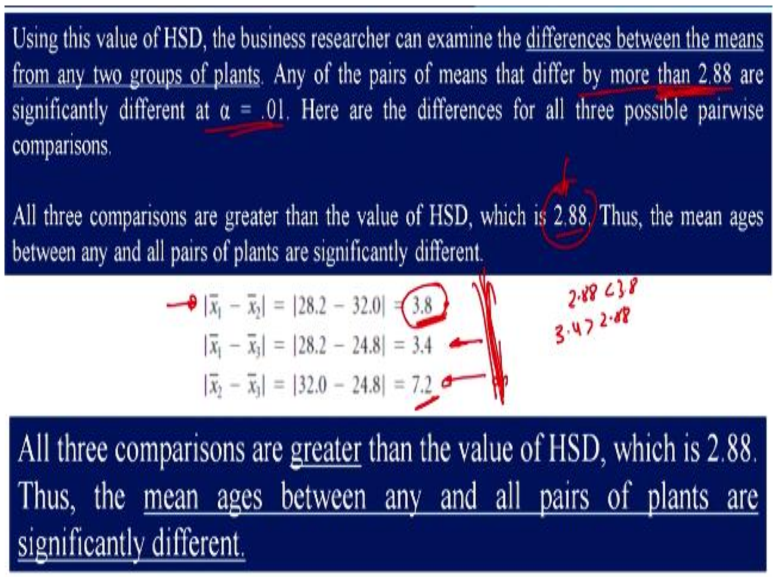

So, we will take up an example and we will try to find answer, right so, this is a question wherein you have got three 3 plants and these are the ages of workers, so we want to find out, is there any significant difference in mean age of these 3 set of workers of three different plants right, so we know that group mean is this for this and for third group; group mean is 24.8, so the point to be remembered here is it is emphasise the same in all these groups, right.

So, because emphasises are equal will use Tukey's HSD test for comparison of means between 1 and 2, 2 and 3 and 1 and 3, right, so we have already solve this question in previous class and this was our solution, right, this was ANOVA table, so we rejected null hypothesis. What was null hypothesis; let the mean age of all these three plants are same, right, so this was mean squared error, just see this is MSE; mean squared error which is required in this formula, just see this.

So, mean squared error is 1.63, 5 is the sample size and this value q alpha is to be obtained from table, right and this is the table at appropriate degrees of freedom, right, so this is c's number of groups, degree of freedom n – c, so this is total number of items, so this 5+5+5, right 15, right, 15 - number of groups, so this 12, so at 3, degrees of freedom of population, number of population is this and at 12 degrees of freedom, so this is 3 and this is 12.

So, this becomes 5.04, right, so this is value over here, so we have done nothing, we have just put the values of MSE and this q alpha value at c and n - c degrees of freedom, so when we calculated HSD, it turns out to be 2.88, right, so remember that HSD is 2.88, right.

**(Refer Slide Time: 14:15)**



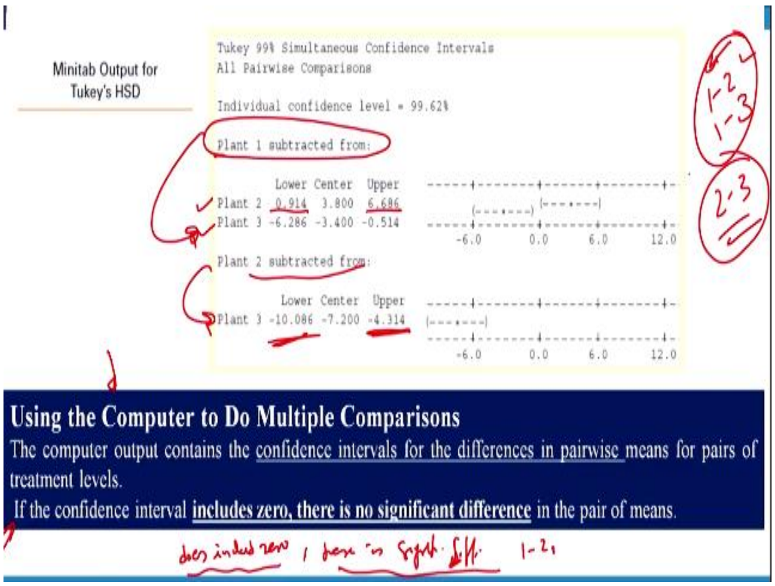

So, using this HSD, the researcher can examine the difference between means from any 2 groups of plants, so any pair of means that is different by more than this are significantly different at this significance level, so you need to find out the means of differences, right. So, let us look at this so, this is the mean of difference between first 2 groups, right, x1 bar – x2 bar, what is x1 bar; mean of first group, right.

What is mean of first group, okay, it is not there, right but you can easily, yeah it is here, right, so let me erase all these, right, so group mean is this 28.2, 32 and 24.8, so group mean between 1 and 2 is this minus this, right, so just see this, we have to take the absolute difference, so this is 3.8, difference between x1 and x3, 3.4 and x2 and x3, 7.2, so now what; you have calculated this HSD.

So, compare this HSD with each of these difference of means, so if you compare let us say 2.88 and 3.2, so 2.88 is < 3.8, so this value is greater than this, it means there is a significant difference between first and second group, so if the difference is more than this, then we will say there is a significant difference between first and second group, what about 3.4 and this; again 3.4 is > 2.88, so there is a significant difference.

Again, there is a significant difference, we will say that there is a significant difference between group 1 and 2, 1 and 3 and 2 and 3, right, so that is our conclusion, right, what we are saying the first thing is there is a significant difference in age of these three plants, this what we did using ANOVA but using Tukey's test, we have said that there is a significant difference between a mean age of plant 1 and plant 2 employees, plant 1 and plant 3 employees and plant 2 and plant 3 employees.

**(Refer Slide Time: 17:09)**




So, this is one way of finding answer, now there is one more way of finding answer and this is known as confidence interval approach, so this is an output from a software, so what we have done here is; so these are confidence intervals, let us look at, this is plant 1 subtracted from so, we are comparing plant 1 with plant 2 and plant 3, first comparison is between plant 1 with plant 2 and plant 3.

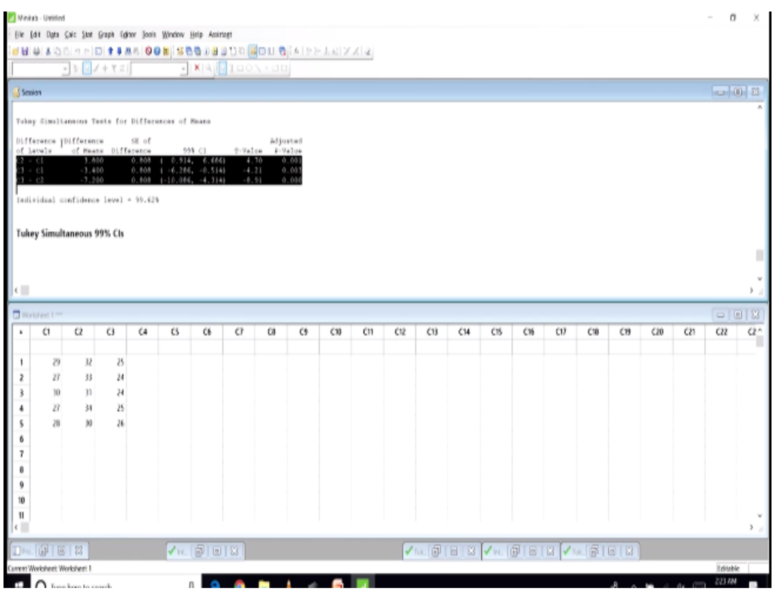
Now, let us look at this, the upper limit is this and lower limit is this so, we can use this computer output for making decision, so we have been given confidence intervals, so we will look at this point, right, if the confidence interval includes 0, there is no significant difference, if confidence interval includes 0, if does not include 0, there is significant difference, is not it so, does it include 0?

When we compare plant 1 with plant 2, plant 1 and plant 2, is there any 0 in this limit; no, so this does not include 0, it means there is significant difference between 1 and 2, what about plant 1 and plant 3, now, does this interval include 0; no, it means there is a significant difference, what about now; now, we have checked plant 1 and plant 1,2 and 1,3, right, which comparison is left now; 13, right, so plant 2 with plant 3.

Is there any 0 in it, does it include 0; this is -4.3142 – this, there is no 0, right, so it does not include 0, it means there is a significant difference between plant 1 and plant 3, so there are 2 ways in which you can find conclusion, right. The first is using just by comparing HSD value with the differences of means, so if those differences are greater than the HSD value, we say that there is significant difference.

Otherwise, there is no significant difference similarly, if the confidence interval does not include 0 then, there is a significant difference, right.

**(Refer Slide Time: 20:32)**




So, let us workout this example using Minitab, so this is our Minitab window, so we will enter data for this particular question, so we have a 29, 27, 30 these are ages of workers of plant 1, right, this for plant 2, all these values for plant 2, similarly for plant 3 so, 25, 24 again, 24, 25 and 26, right.

So, we will go to stat ANOVA one way, so first of all responses in separate column right, so we have just got these 3 columns; C1, C2, C3, select all of them, will go to options, we will select confidence level, so we will see what was the confidence level right, so in this question the alpha was 99%, right, so this is 99%, right and it would be a one tailed test as we have already seen it is always a lower tailed test, so we will click, okay.

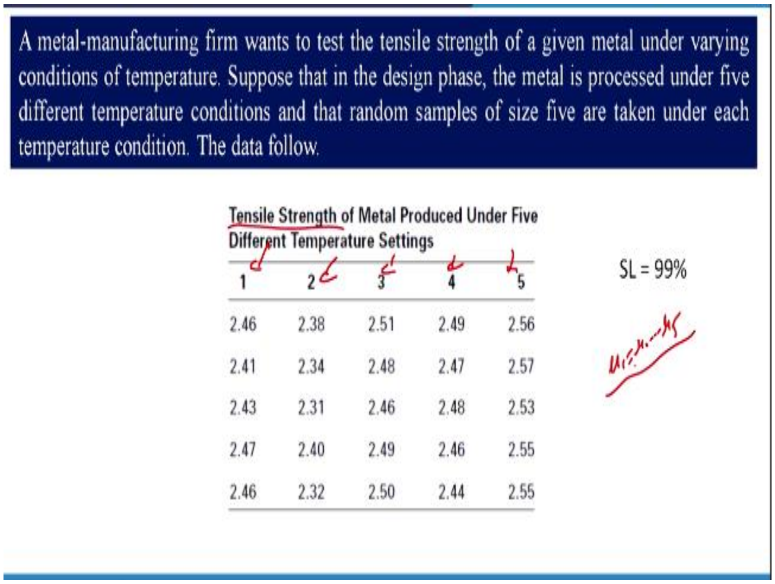
Then, we will go to comparisons, right, so this is error of; error rate of comparison is 1, right, so it is 99, so 1, right, 95, 5 right, so Tukey’s test, right and then we will click okay and okay, right, so we will not look at this plots, we will just look at the table, so first of all look at this ANOVA table, right, so p is this right, p0 right, it means we will reject null hypothesis, right, this one conclusion, right.

The next one is; if you look at this in fact, you will get confidence interval in such a way that you will not have 0 in between confidence intervals in fact, we can redo it, just go to stat ANOVA one way, we look at comparison, so let us look at this test, we can click it over here test and yeah let us look at this, yeah these are confidence intervals now right. So, look at C1 and C2, it does not include 0, so there is a significant difference, right.

C1 C3 it does not include 0, so there is a significant difference, third confidence interval; again it does not include 0, so there is a significant difference, in fact interestingly, you can have in fact p values also, right, so alpha was 0.01, right and in this case all these p values are $< 0.01$, so we will reject null hypothesis, right, so we will say that and there is a significant difference between C1 and C2, C1 and C3 and C2 and C3, right.

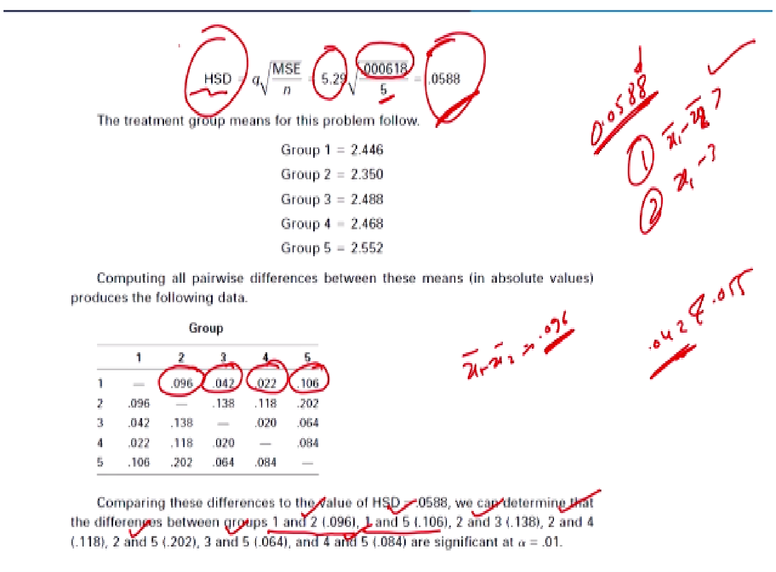
So, this is how you can find out answer to a question on Tukey’s test, right, so we will solve one more example on Tukey’s test.

**(Refer Slide Time: 24:51)**




So, this is the question; a metal manufacturing firm wants to test the tensile strength of a given metal under varying conditions of temperature, so we have to find out is there any significant difference in tensile strength under different varying conditions and these are 5 different varying conditions or these are 5 different conditions of temperature, so null hypothesis is that there is no significant difference, right, Mu 1 to Mu 5, right.

**(Refer Slide Time: 25:40)**



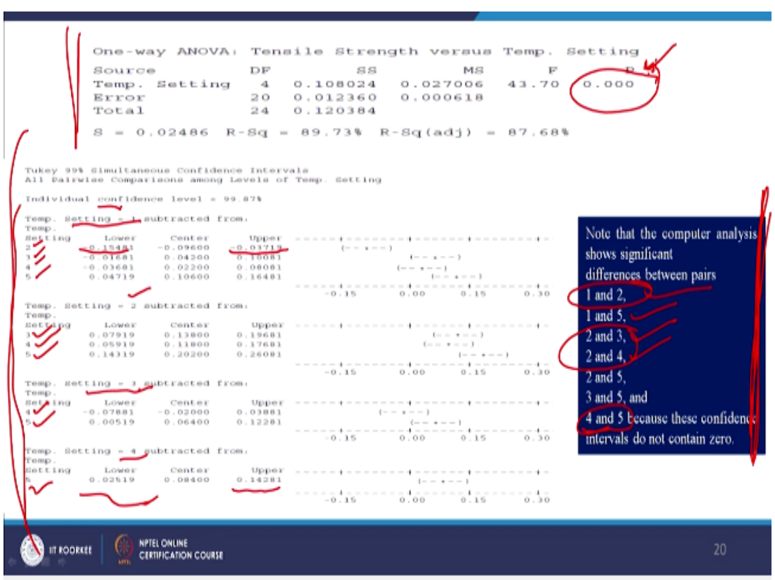

And there is a significant difference would be alternative hypothesis, right, so let us look at find out HSD value, so for this question you need to calculate first mean of squared error, so this would be the mean of squared error, it is 0.00061 divide by sample size and this is your q alpha, value, 5.1, so at the end of the day this is your HSD value, 0.0588, so what we concluded in previous example that if this value is small than; is smaller than all those differences, it means there is a significant difference amongst all those groups.

In other words, if the mean difference of x1 and x2 is > this, then there is a significant difference so, first of all what we need; we need is we want to find out differences, right, x1, x2 then second would be x1, x3, right, so x1 x2 difference is this, is not it then, x1 x3, x1 x4, x4 x5, so these are 5 differences, right, similarly x2 3, 2 4, 2 5 then 3 4, 3 5, 4 5, right, so these are the differences you can easily find out.

Now, let us compared the first difference, right, x1 – x2, right, this difference is 0.096, okay, what about this; so this value is greater than this value, so it means there is a significant difference between group 1 and 2, okay so, there is a significant difference between group 1 and 2, 1 and 5, 1 and 2, what about 1 and 3, if you look at these 2, right, 0.042 and 0.053, so this is smaller than right.

So, there is no significant difference between 1 and 3, so the significant difference between 1 2, 1 5, 2 3, 2 4, 2 5, 3 5 and 4 5, now this is the conclusion we have got from HSD value.

**(Refer Slide Time: 28:22)**



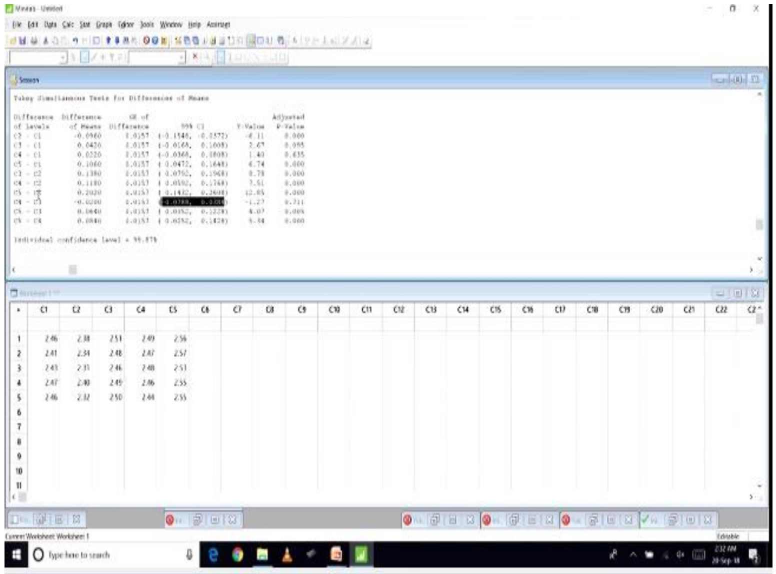

The same conclusion can be find out; can be found out using confidence interval approach, right, so this confidence intervals of different groups, right, so let us look at this so, what is our; when do we say that there is a significant difference; if last interval does not include 0, there is a significant difference, right, so if you look at this particular question, there is a significant difference between 1 and 2.

So, this is 1 right, this is setting 1 and 2, right, this is 1 and 2, 1 and 3, 1 and 4, 1 and 5, so between 1 and 2, let us look at this, this and this does this interval contain 0, does it include 0, it does not include 0, it means there is a significant difference similarly, 1 and 5, let us look at 1 and 2, so this is -0.12, sorry this + 0.1 in the upper limit and lower limit is this, it means there is 0, so there is no significant difference between 1 and 2.

But there is a significant difference between 1 and 5, 2 and 3, 2 and 4, you can see 2 and 2, 3 2, 4 and 2, 2 5, so this is 2, 2 3, 2 4 and 2 5, now this is 3 and 4, 3 and 5, 4 and 5, so there is a difference, right, 4 and 5, so 4 and 5, right, so does this include, so this interval does not include 0, it means there is a significant difference, right, so this is the output, this computer output, output from any software and this is ANOVA table as well.

So, from looking at here, you can say that there is a significant difference amongst; so by looking at this we are rejecting null hypothesis, right and there is a significant difference amongst one of those 5 groups, right and from confidence interval approach, we can find out in which groups there is significant difference, right.

**(Refer Slide Time: 30:58)**



So, let us work out this example using Minitab, so this is our Minitab output, so we will delete the previous output, so let us look at these values is 2.46, 2.41, 2.43, 2.47, 2.46, 2.38, 2.34, 2.31, 2.40, 2.32, just keep in mind that the sample size of each of these groups is same, it is 2.46, 2.49, 2.50, 2.49, 2.47, 2.48, 2.46, 2.44 and the last group; fifth group is 2.56, 2.57, 2.53, 2.55, 2.55 so let us see the output, right.

So, ANOVA one way, this is C1 to C5, select all these, click over here, this is in separate column, yes that is correct, so we will at what test significance level we were testing that hypothesis; just a minute, so it was at 99% significance level, right, so this 99% significance level okay, so will have a right, Tukey test, everything is fine, so we will look at now this confidence interval, right.

In fact, we can look at even p value as well, right so, just by looking at p value itself we can say that there is a significant difference between first third this, C5, C1, C3, C2, C4, C2, right and C3, C5, C4, C5, again look at confidence interval, does this interval include 0, it does not include 0, it means there is a significant difference between C1 and C2, does this include 0, this confidence interval does not include 0.

So, there is a significant difference between C4 and C5, if you look at this let us say this, okay now this include 0, so there is no significant difference between C3 and C4, so this is how you can use Tukey’s test to know the significant difference between groups. So, with this let me stop here, in next class, we will see a method wherein we will have an example were the sample sizes are different in each of the groups for which you would be taking up for solution, so thank you very much.