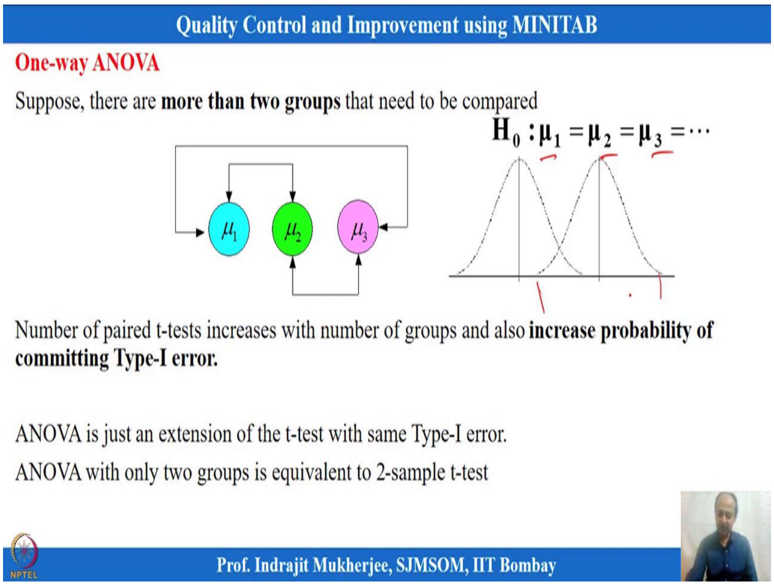
Quality Control and Improvement with MINITAB
Prof. Indrajit Mukherjee
Shailesh J. Mehta School of Management
Indian Institute of Technology, Bombay

Lecture - 21
One-way ANOVA

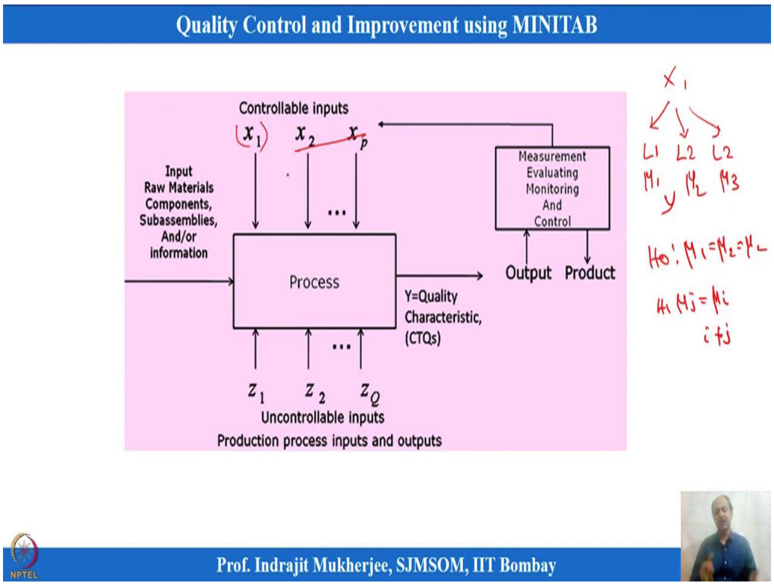
Hello and welcome to Session 21 of our course on Quality Control and Improvement with MINITAB. I am Professor Indrajit Mukherjee from Shailesh J. Mehta School of Management, IIT Bombay. So, in this session we will see some examples of doing One-way analysis of variance.

(Refer Slide Time: 00:39)



So, last time in our course we were are trying to understand what is analysis of variance and why we are using that?

(Refer Slide Time: 00:52)



So, let us assume that there is one factor and this diagrammatically explains what we are trying to do. So, there is only one factor. I can ignore the other factors over here. And there is one factor X_1 and I want to check whether at different conditions of X_1, whether that mean response (y) is changing over here or not. So, here the average response we can think of μ_1, μ_2, and μ_3.

So, we want to check whether everywhere this average value that we are getting is same or not or whether there are any 2 levels such that $\mu_i = \mu_j$. Alternate hypothesis is $\mu_i \neq \mu_j$. So, there are at least 2 levels where the average is quite significantly different.

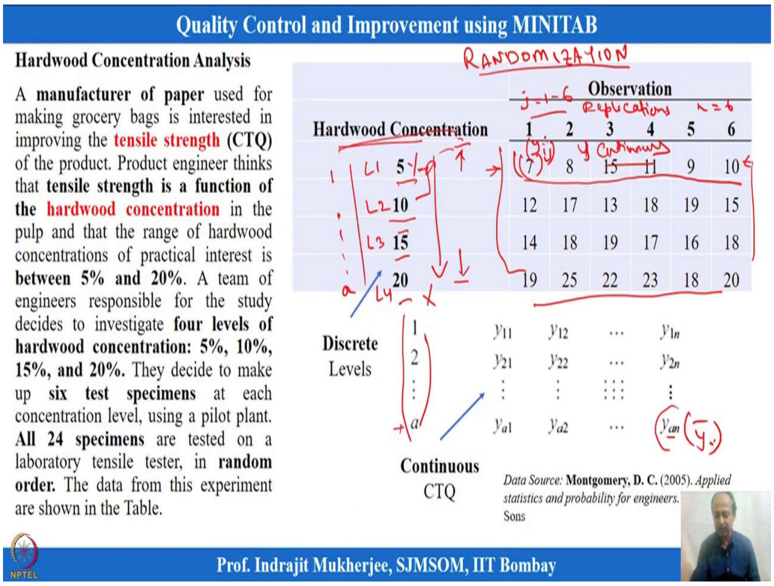
So, this is developed when we have one factor. So, this is not the scenario in most of the design of experiments, but this is a most favorable scenario we can expect when I have one factor and I want to check the optimal levels and find out what is the optimal levels which will optimize basically Y or CTQs.

And when there is only 2 levels like that, we have suggested 2 sample t-test for that or paired t-test in certain scenarios like that ok. So, if you have a factor controllable factor which is in your control as an experimenter, and you have more than 2 levels then type 1 error can be controlled. So, the type 1 error will not increase if I apply analysis of variance instead of 2 sample t test. So, t-test for more than two levels is not recommended, what is recommended is analysis of variance proposed by Ronald Fisher approximately around 1921.

This was very popular when it was proposed and people accepted this one and still people are using analysis of variance in design of experiments. So, we are at the improvement phase what we are discussing now, analysis of variance, 2 sample t-test all are in improvement phase.

We want to check when we have done improvements whether it is effective or not so, statistically whether they are different or not. Here also we are trying to check whether the factor is significant or not which influences y or not and what should be the level of x that will optimize the y over here. For this analysis of variance is suggested. So, here you can see that we are comparing different means and the variance information is used over here. Variance information is used over here to see the mean difference basically. So, that is why it is known as analysis of variance.

(Refer Slide Time: 03:42)



Then we I explained how people are doing experimentation with one single factor. So, here the factor is hardwood concentration and it has different levels 1, 2 and this is 3 and this is 4 like that and this are known as replicates. So n equals to 6 over here; that means, at 5 percent hardwood concentration experiment was done with different samples and 6 different samples over here and the 6 observation with 5 percent concentration is given over here. Similarly, at 10 percent. But this experiment was done randomly.

So, randomization was another important concept that was introduced. And why it is required we will understand afterwards, but at present we should know that randomization is basically the concept that you have to implement when we are going for experimentation. Every experiment has to be randomized, and how do we randomize? We select any of these levels over here and we select any of the samples and combination of that the results what tensile strength that is generated over here this is known as y_{ij} let us say observation number i over here and j varies from number of levels over here is what we have defined. j varies from 1 to 6 like that. So, this is the j variation over. So, we have to understand which is a factor, how many levels we want to experiment. This you have to freeze before doing experimentation. How this range of the factor or how this level of the factors is selected? It is based on engineering judgment that this is the variation or the process can go up to these extremes. So, factor's range feasibility we have to check and based on which we have to select the levels and also there should be gap between the levels, it should not be very close.

You can see books how the levels are selected. So, this is one example where I want to maximize tensile strength and hardwood concentration experiment was done, randomization was implemented over here, 6 replicates are taken over here and first experiment maybe with 5 percent, second maybe with 10 percent, then 20 percent, then 15 percent like this and we get total 24 observations. So, 6 multiplied by 4 levels over here, 24 observations reading we have got.

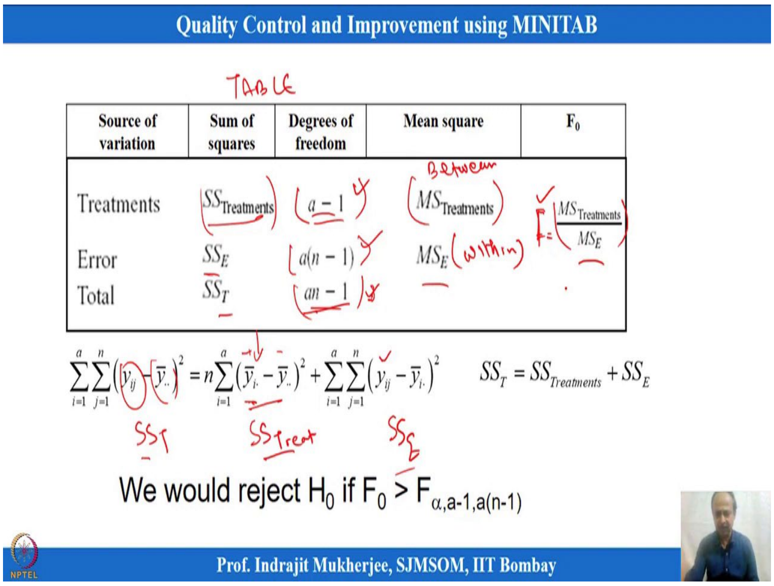
And we want to analyze this data and try to figure out at what level we should freeze the hardwood concentration if this is the only factor and then we can also see that when I change the levels whether it is impacting the mean value of the response.

Assumptions over here, is that in analysis of variance the factors have discrete levels. These are the discrete levels that we are experimenting over here and the y that we are getting over here is basically continuous variable.

One way means one factor at different levels which is more than 2.So, I am changing the levels and this is in my control basically. The statistical model that is used here is known as fixed effect model. So, for that model we cannot generalize. We can say for 5, 10, 15, 20 like that these are the levels based on which we are making a judgment and we cannot generalize it for any values between 5 to 20. So, that is when we want to do that that is random effect model basically.

So, here what we are doing is that fixed effect model that is these are the levels discrete levels, and this is the outcomes of the experimentation and from here we want to determine which is the best level.

(Refer Slide Time: 08:25)



Then what I told is that, there is a ANOVA table. So, this is the table that we will get and in this case what we observe is that some $SS_{Treatment}$ calculation is done. $SS_{Treatment}$ is basically the variation of each individual observation average. So, we will get some average over here at a particular level, and the from the overall average that is the grand average that we will get $\bar{y}_{..}$.

So, that variation which we capture is known as $SS_{Treatment}$. So, this is represented over here formula which I have not told earlier. SS_{Total} is the overall variation and represented as:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

i.e. individual observations minus overall grand observation. Then what we have is that error variation SS_{Error} over here. So, because not only factor influences the overall variation of the process, there can be other X which we do not know like that. So, there will be some error in the estimation over here. So, that is known as SS_{Error} which is calculated as:

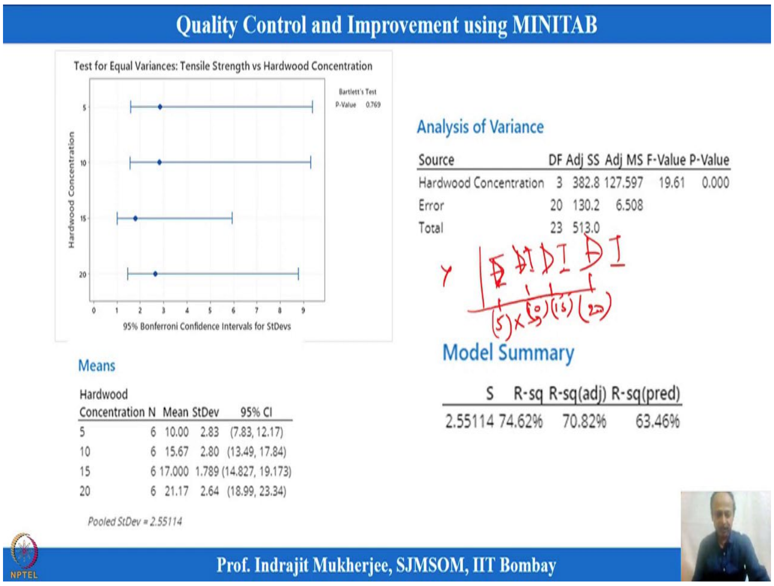
$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

Then there are degrees of freedom. If I have a levels, then $a-1$ is the degree of freedom. If I have total 24 observations in experimentation with replicates then in that case $an-1$ is the degree of freedom for this. And when you divide $SS_{Treatment}$ by $a-1$, I get a mean square treatment.

So, $MS_{Treatment}$ is between variation and MS_E is the within variation. Fischer recommended that you calculate a statistic which is $\frac{MS_{Treatment}}{MS_E}$ that give you F-value which follows F distribution basically and these F values can be compared with tabulated values.

If the F values are higher than F tabulated value then in that case we can expect p values to be less than 0.05 like that and we will go by p values in MINITAB analysis.

(Refer Slide Time: 11:54)



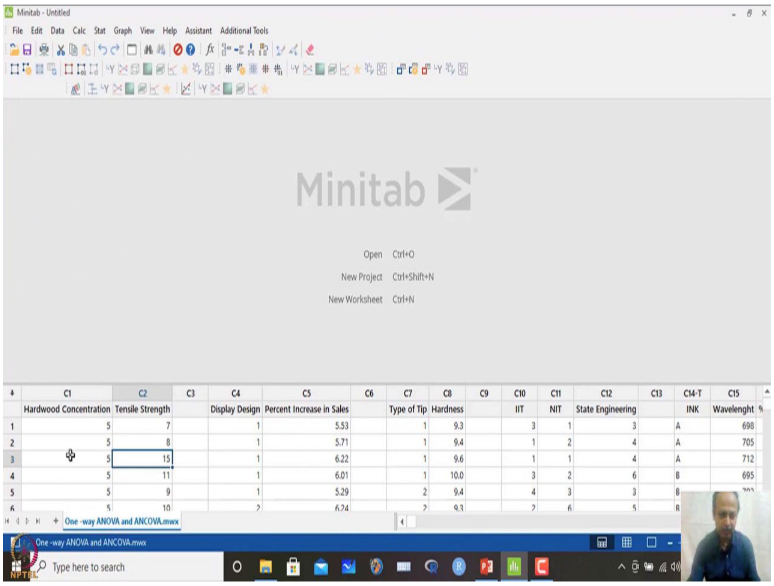
So, let us do the experimentation now with using MINITAB and for this what is required over here is that, some assumptions has to be verified initially. One of the assumptions which is required is that whether the variance at different level is same or not. So, if I plot X as different levels over here 5 percent, 10 percent, 15 percent and 20 percent like that versus Y. So, in these case we expect that there will be variation because the experiment if I repeat n number of observations I have over here at 5 percent that is 6 observation I have. And this can vary, this will vary. Basically we cannot get single value.

So, similarly 2nd value will also have some variation, 3rd value will also have some variations like this and we want to check whether the standard deviation over here or variance that we are estimating over here and variance that we are getting over here are all same or all different. Because based on that our analysis will change. So, in this case like two sample variance we are testing is there. Here also if there is more than two sample variations over here and we can compare that whether the all variance are same or whether any two variance are different like that.

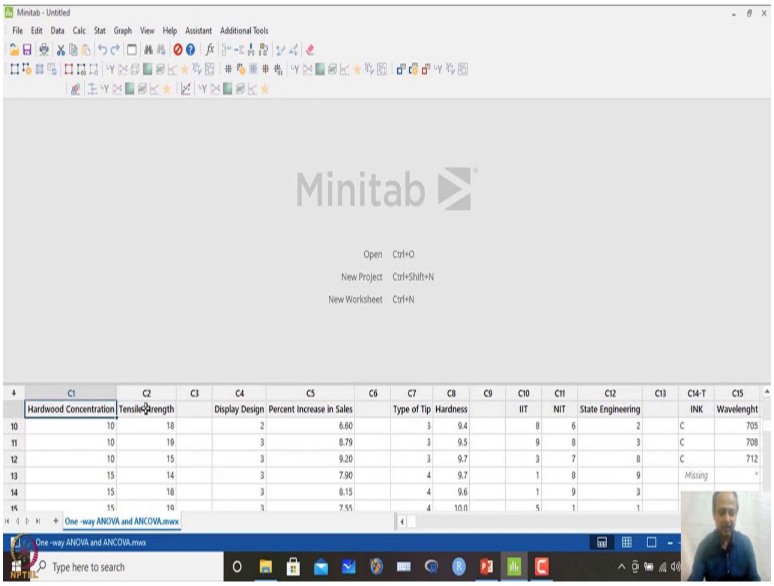
So, for that and underlying assumptions has to be made over here that is whether the values in each group over here follows normal distribution or not each of them follows normal distribution or not based on that test will also differ. So, we have to first test that group wise whether they are normally distributed or not.

Individual at 5 percent, 10 percent, 15, 20 and if they are assumed to be normal distributed in that case what we can do is that, we will assume that one and go ahead with the test and if it is not true then we will go ahead with the different test which is known as Welch's test and that is also possible in MINITAB.

(Refer Slide Time: 13:54)



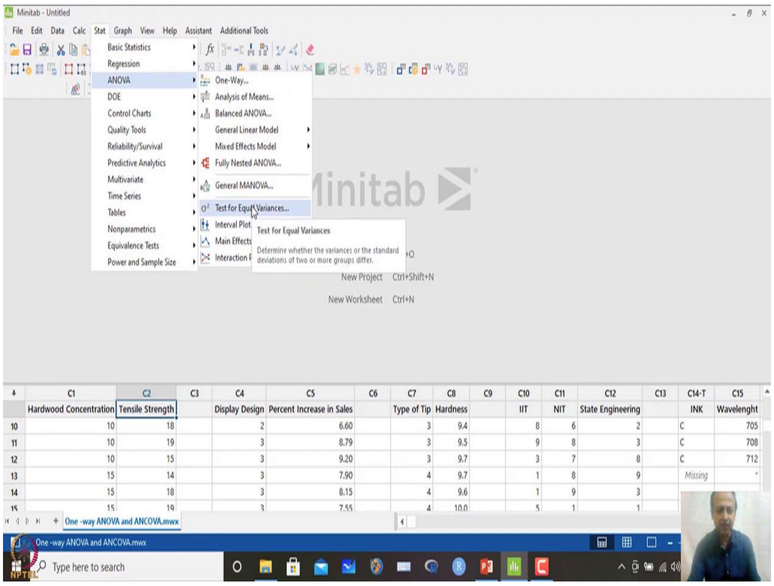
(Refer Slide Time: 14:02)



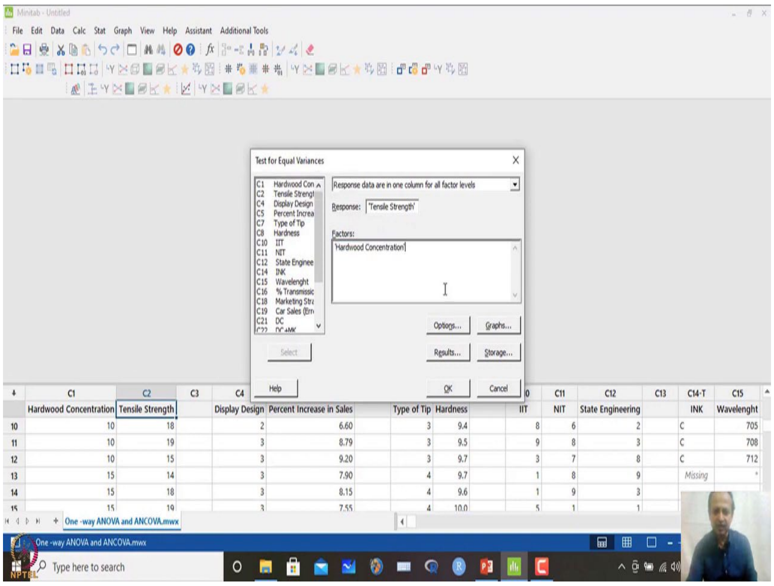
We are trying to analysis this dataset and which is the hardwood concentration and then we will see what to do and how to analyze the ANOVA and how to interpret the

ANOVA analysis. So, here hardwood concentration is changed and data is given over here and tensile strength data is given over here.

(Refer Slide Time: 14:07)



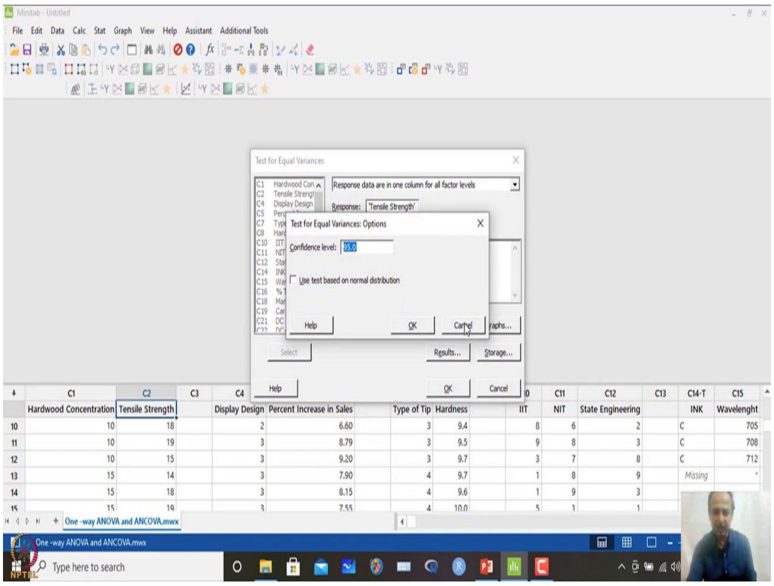
(Refer Slide Time: 14:17)



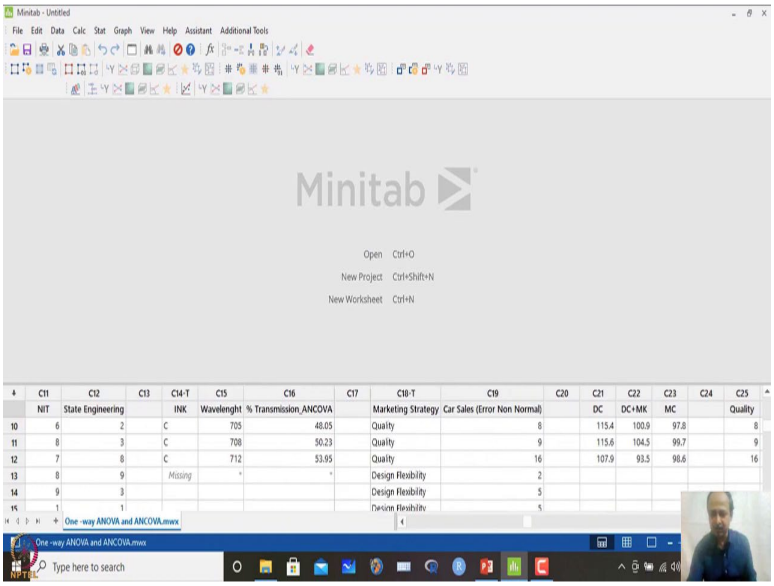
So, what we will do is that we will go do *stat* and in ANOVA analysis there is a option of test of equal variance over here. So, what we will do? First we will test whether the variance condition that is required which is satisfied or not over here. So, then what I will do is that we will just see that each factors are in same columns.

So, in this case what is the response variable over here? The response data are in same column. We will just highlights C2 over here and factor that we have to give here is hardwood concentration over here.

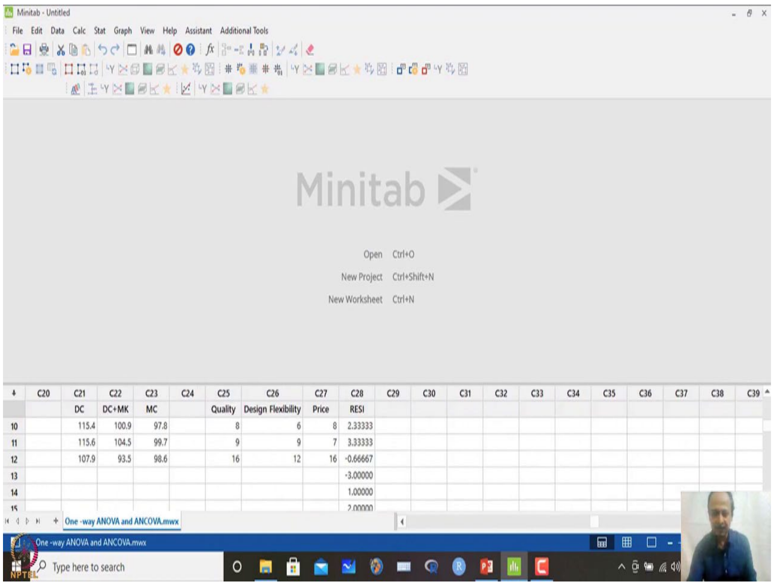
(Refer Slide Time: 14:45)



(Refer Slide Time: 14:56)

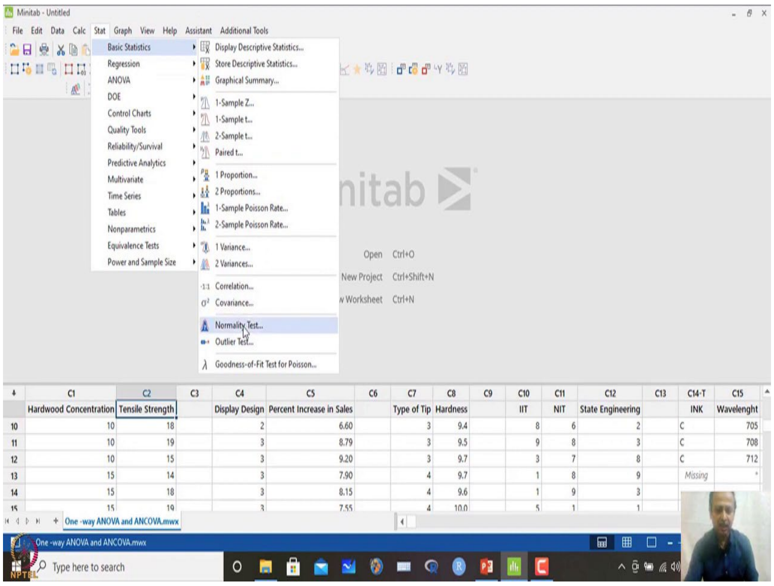


(Refer Slide Time: 14:58).



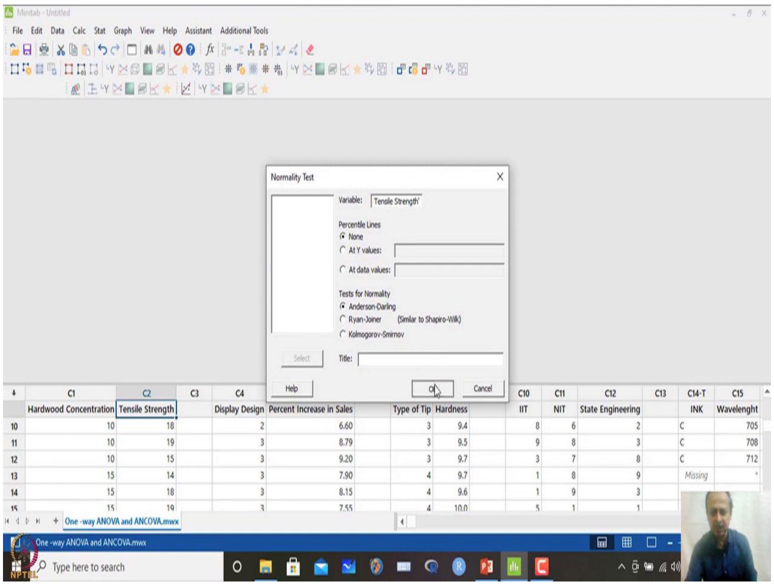
Then you go to options over here and you test for normality. So, this check has to be done over here. So, what do you have to do is that, you have to just separate the values for 10 percent and 15 percent, 20 percent, 5 percent like that. And if you can differentiate that one and then check. So, I have to segregate that one.

(Refer Slide Time: 15:09)

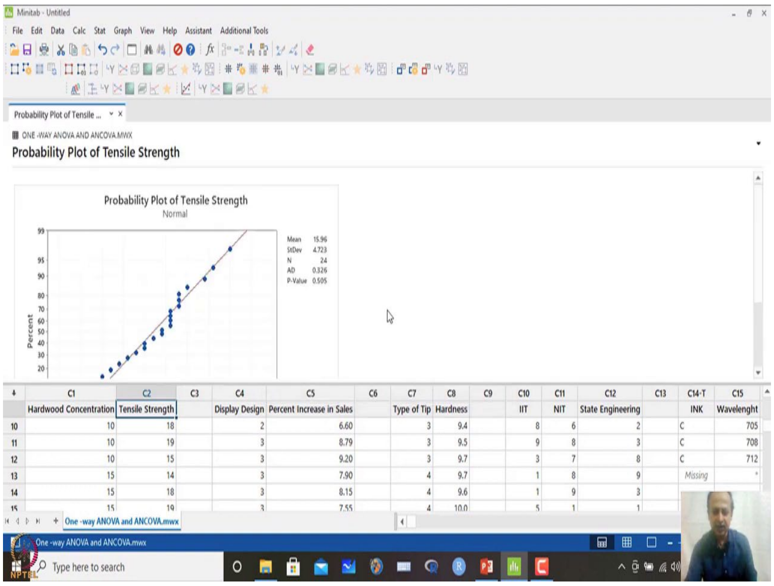


So, I am doing a rough approximation over here and trying to check whether the overall values is normal. So, I will do the basic statistics, but what do you have to do is that, 5 percent you have to segregate, the dataset 5 percent, 10 percent and individually you have to see whether it follows normal or not. I am taking a overall test of tensile strength over here which I am doing normality test. So, I will take the tensile strength and want to check Anderson Darling test.

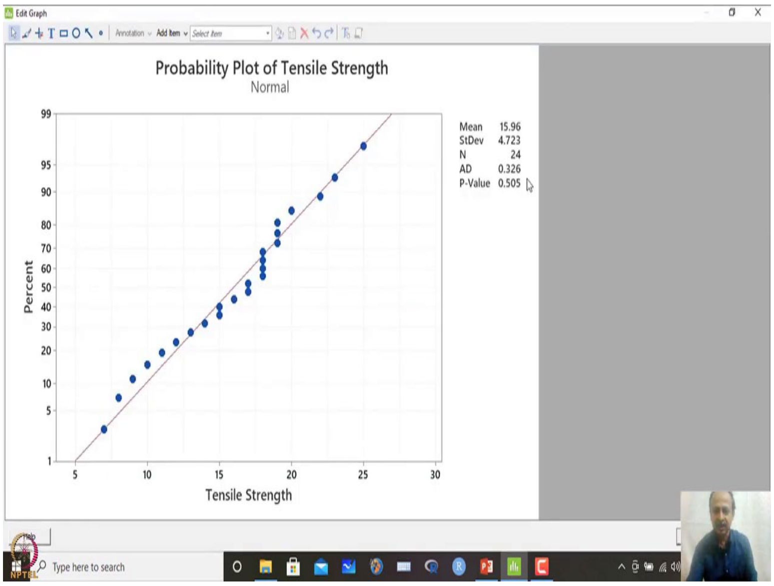
(Refer Slide Time: 15:22)



(Refer Slide Time: 15:27)

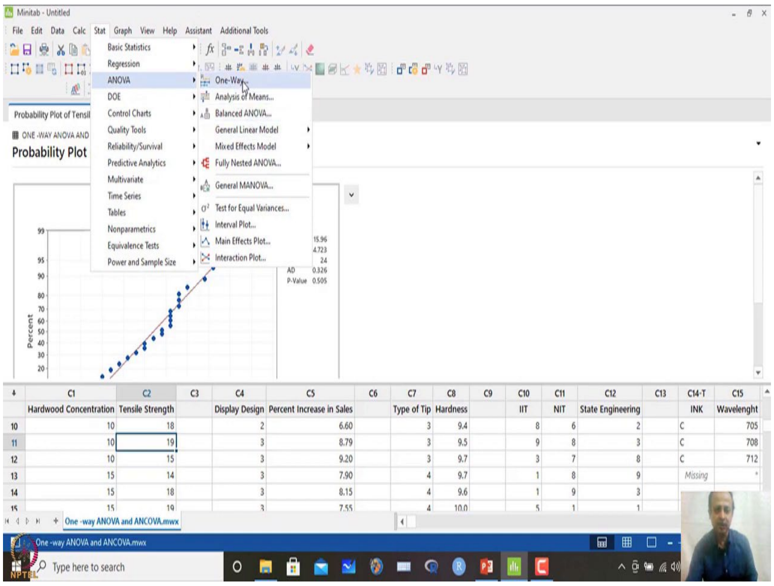


(Refer Slide Time: 15:33)

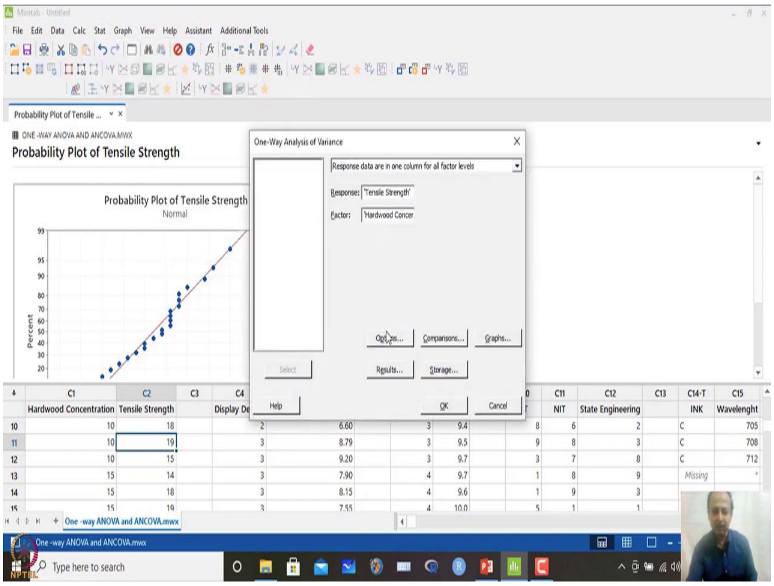


So, if you click this one what you get is that, you get values approximately like this P value is 0.5, but group wise we have to do like that, but I am doing a overall analysis over here and in this case it shows that mostly we expect that this is a normal distribution data because p value is more than 0.05.

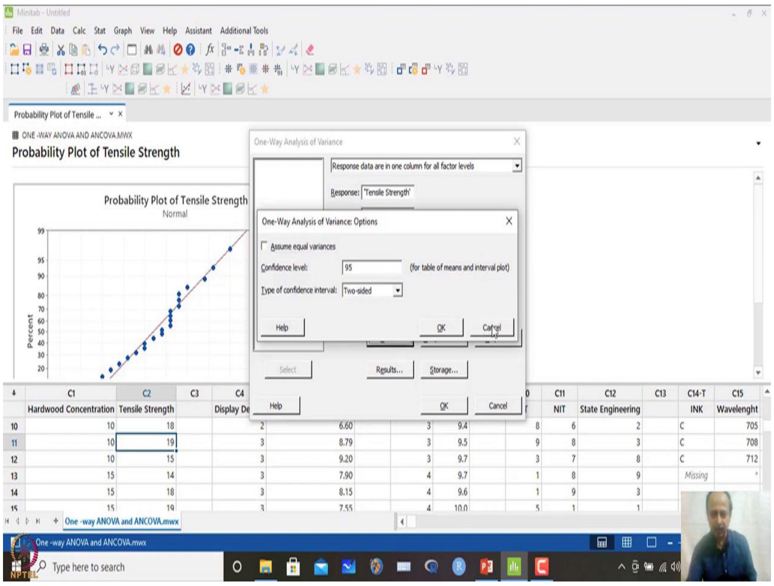
(Refer Slide Time: 15:51)



(Refer Slide Time: 16:00)

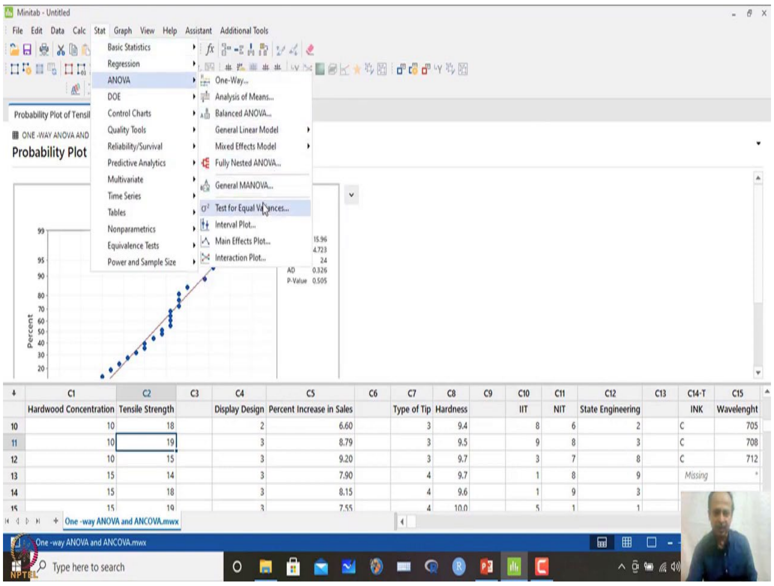


(Refer Slide Time: 16:02)

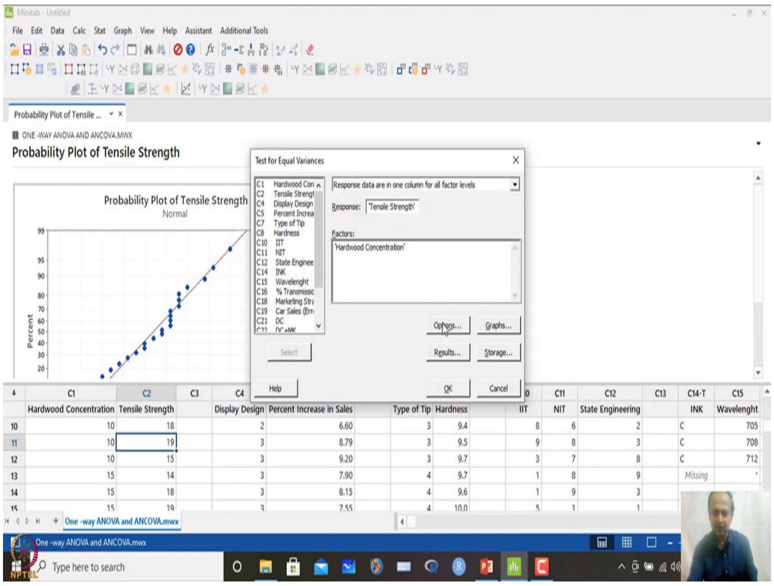


So, assumptions of normality I am considering over here where while checking the variability, variance whether it is same or not then I go back to ANOVA and I will do one-way analysis of variance. And then in the options I will write that assume equal variance, assume normality over here.

(Refer Slide Time: 16:12)

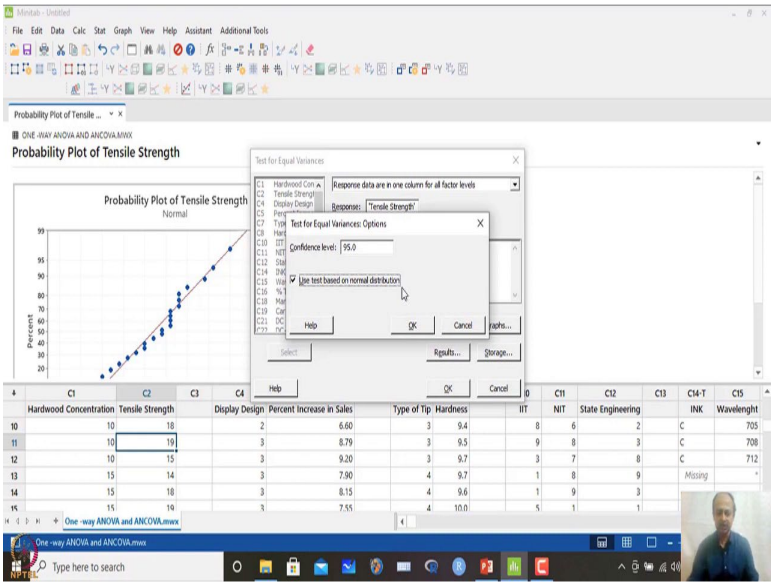


(Refer Slide Time: 16:17)

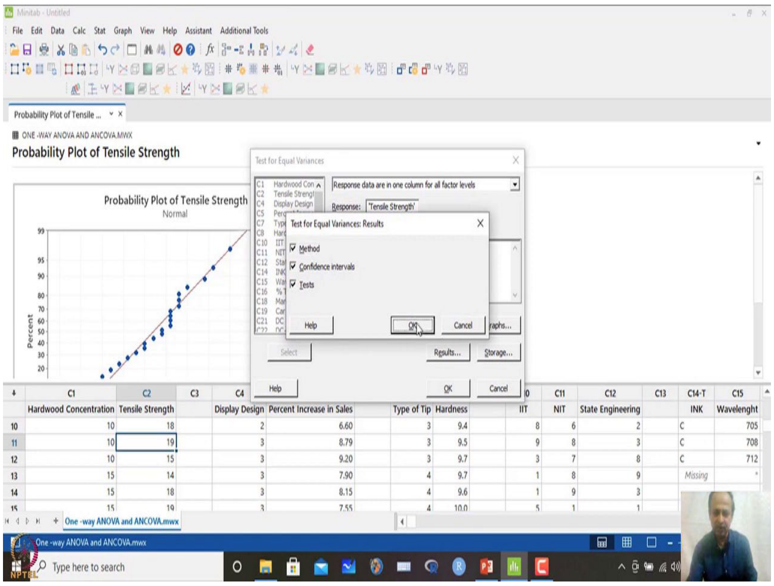


Sorry this is not the test that I have to do. I will select *stat, ANOVA analysis*, *test of equal variance*. So, over here what we have to do is that response data are in one column. So, tensile strength and hardwood concentration we have given.

(Refer Slide Time: 16:25)

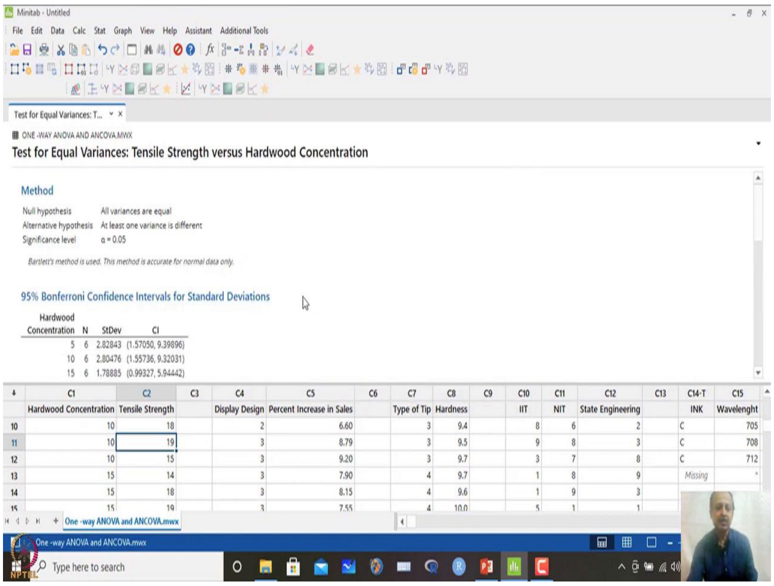


(Refer Slide Time: 16:37)

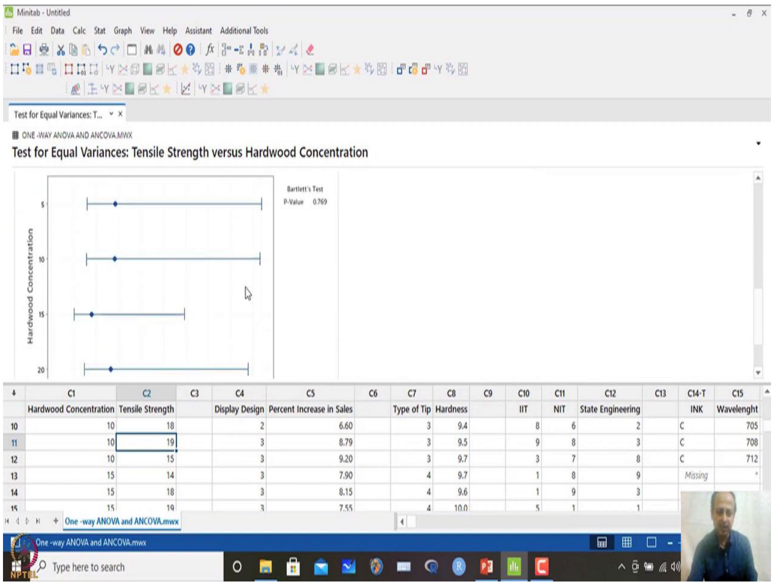


In options we select use test for normality distribution assuming normality distribution over here. I do not change the confidence level over here. So, confidence will remain same. And results what we want to see all things we want to see over here let us say and I click ok.

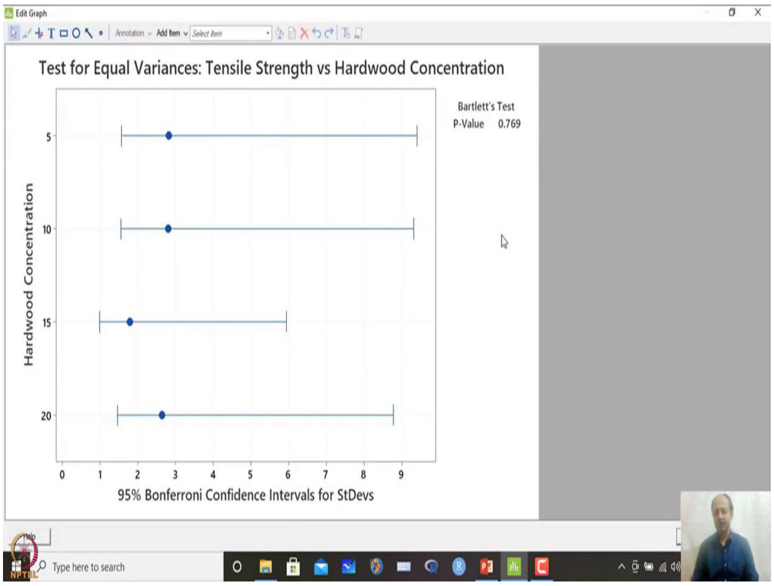
(Refer Slide Time: 16:42)



(Refer Slide Time: 16:43)

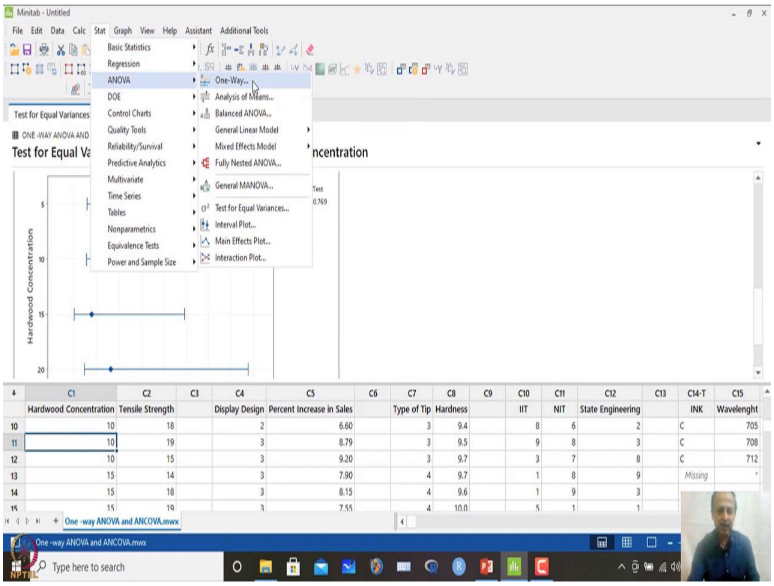


(Refer Slide Time: 16:47)



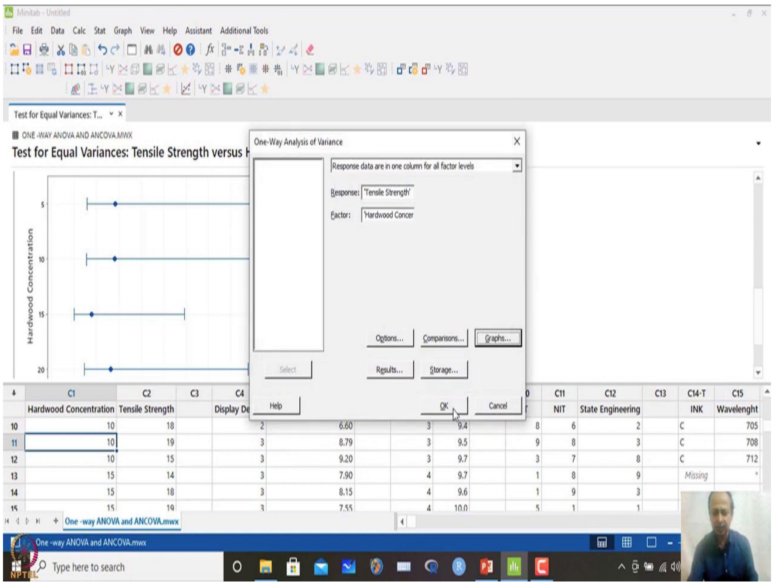
And what will happen is that it will give me some results over here which is the Bartlett test that you will observe over here, because normality assumptions is taken over here. So, it is most suitable test that is possible to do and statistician recommended this one.

(Refer Slide Time: 17:22)

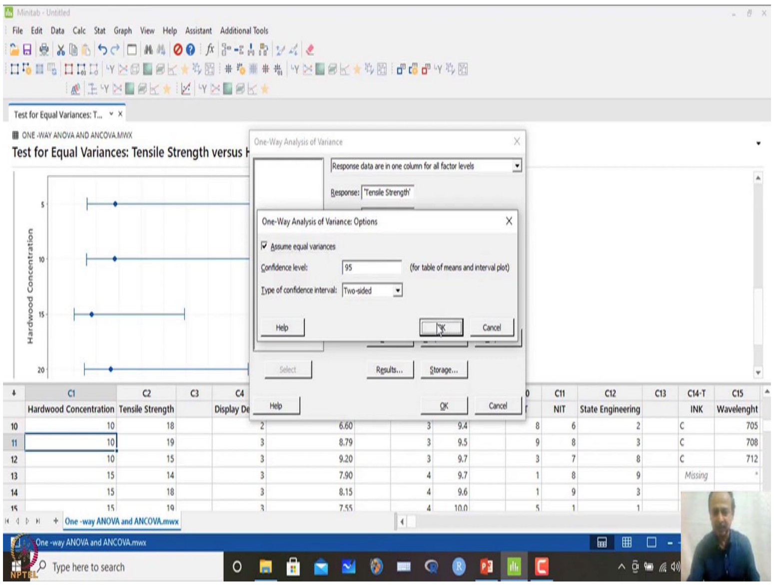


And see the p -Value. p -Value is more than 0.205 indicates that all variances are same. So, overall there is no statistical difference in the variance like that. So, when this test is completed. So, equal variance assumptions, is checked. So, then what I will do is that, I will go to *ANOVA* analysis *one-way ANOVA* analysis.

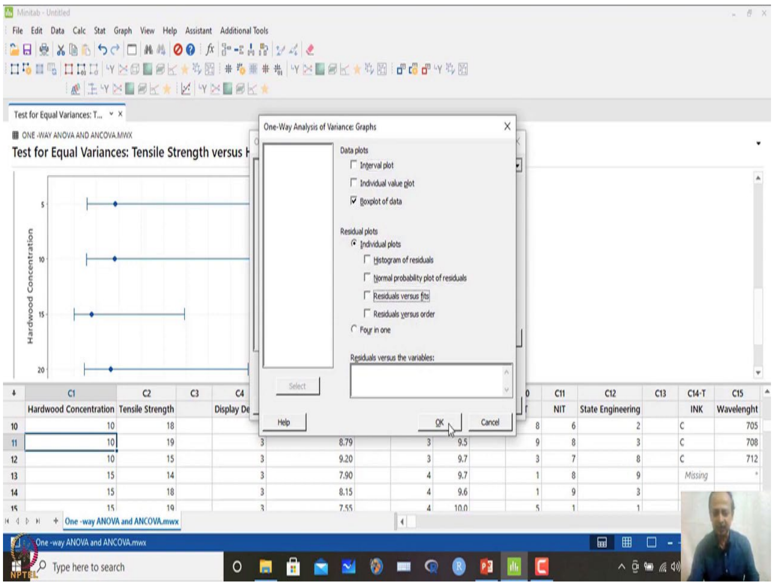
(Refer Slide Time: 17:26)



(Refer Slide Time: 17:29)



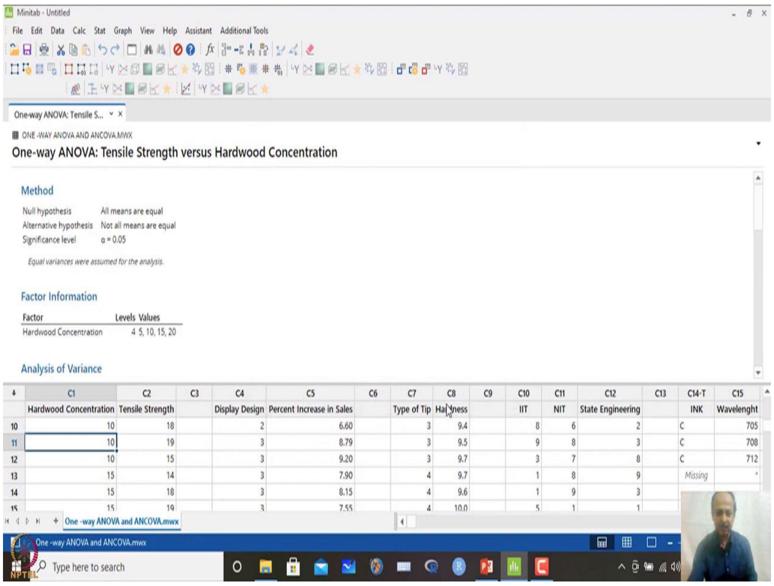
(Refer Slide Time: 17:39)



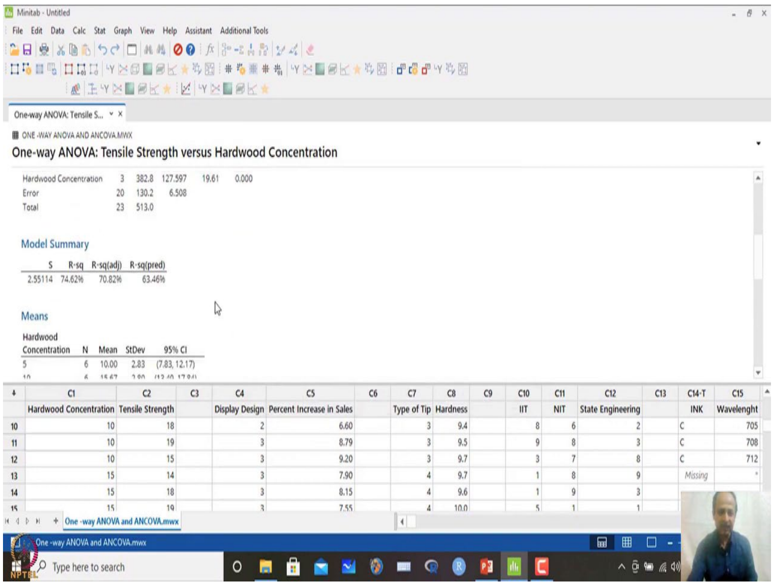
In options I will click assume equal variance; that is because we have already checked that one. Then I will click ok and then I will click in the graph what you can see is box plot we can see over here and also some assumption has to be checked that we will see afterwards.

So, let us do this and let us try to figure out ANOVA analysis says when equal variance condition, normal distribution condition and group wise normality distribution also holds.

(Refer Slide Time: 18:05)

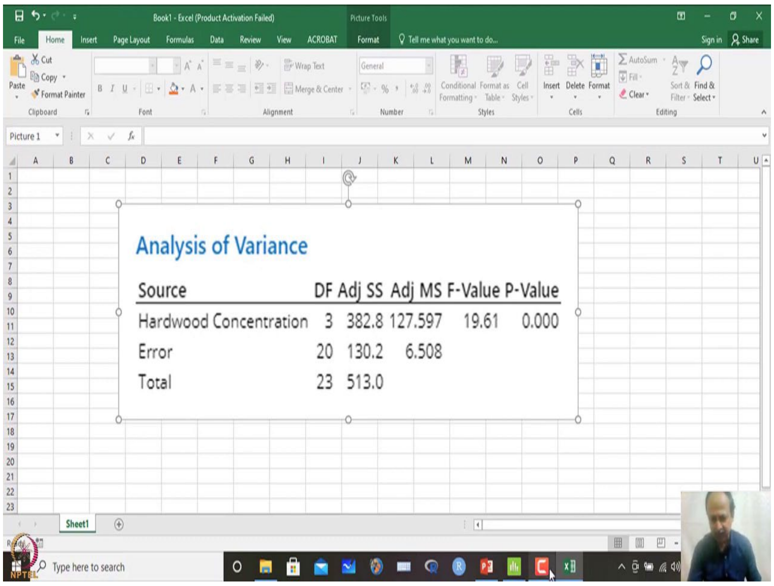


(Refer Slide Time: 18:07)



So, in this case if I click ok what will happen is that, I will get the ANOVA analysis, which is shown over here. So, this can be copied as a picture like that and we can paste it in excel to enlarge the views like that and let us try to see what the results indicate.

(Refer Slide Time: 18:30)



What you see is that source of variation is this SS hardwood concentration. So, when I change the hardwood concentration what is the change in variance that is estimated over here and it says adjusted SS. Formula remains same what we have shown in our slides and MINITAB will say adjusted SS.

So, this is nothing but sum of square variations due to treatment basically. So, this is showing you due to treatment what is the variation then degree of freedom is 3 because we have 4 levels. So, 4 minus 1 is 3 over here and mean square error is SS divided by degree of freedom that is 127 over here. Total degree of freedom is $24-1=23$, over here and sum of square of total is calculated. Now, if you subtract sum of square total and then from that you subtract 382 you will get 130.2 over here. Then mean square treatment is 127 how it is? 382 divided by 3 that is 127 over here.

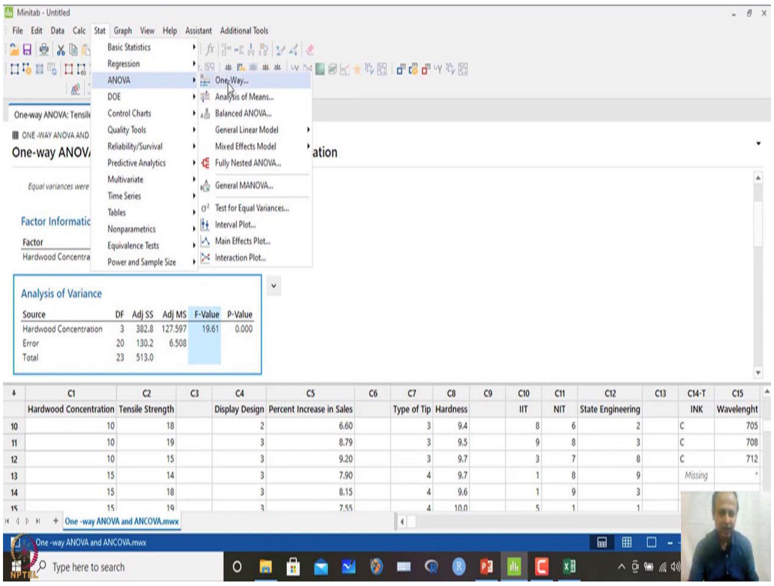
And mean square error what we can calculate is 130 by 20 that will give you 6.5 over here. Then F-value, how we are deriving this one? Mean square treatment divided by mean square error over here. Mean square error gives you an estimation of standard deviation of the process basically or it gives you an estimate $\hat{\sigma}$ of the y and also gives you an estimate of error.

Error variance basically error variance over here. So, over here what you see, F-value is coming out to be 127 by 6.5 this value when you divide it is 19.61 which is very high values of F and it is expected that if F is quite high on the higher side what we can expect is that p value should be going down.

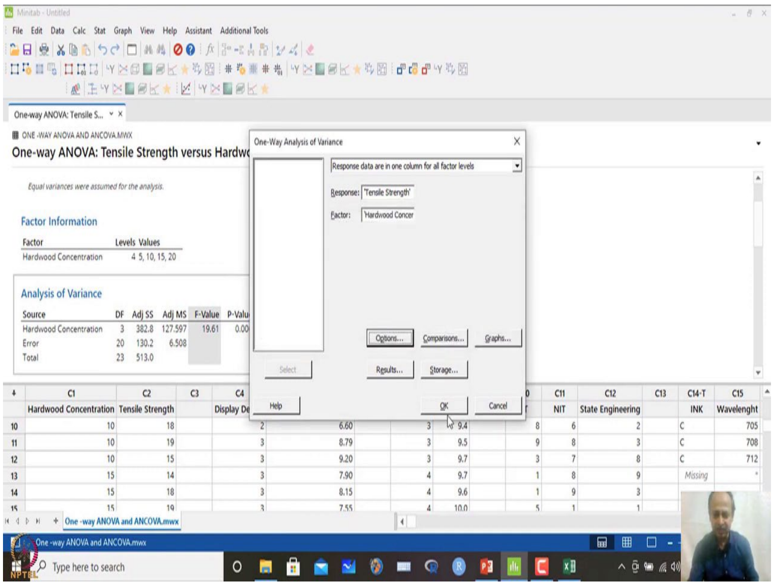
So, highly significant what we are seeing is that, there is at least two levels where when I change the level from one level to the other level basically significant difference exist between the average response that we are getting; that means, this variable is important and is influencing the change in mean of the response CTQs like that.

So, this factor can be considered for further experimentation in future, but this is based on certain assumptions like normality distributions. So, that needs to be checked, but if you are not doing this and if you are assuming that variance is different.

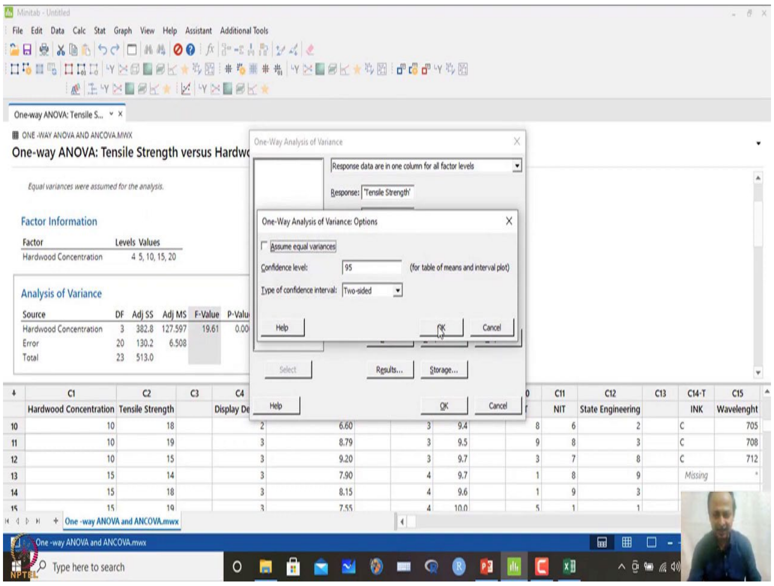
(Refer Slide Time: 21:17)



(Refer Slide Time: 21:22)

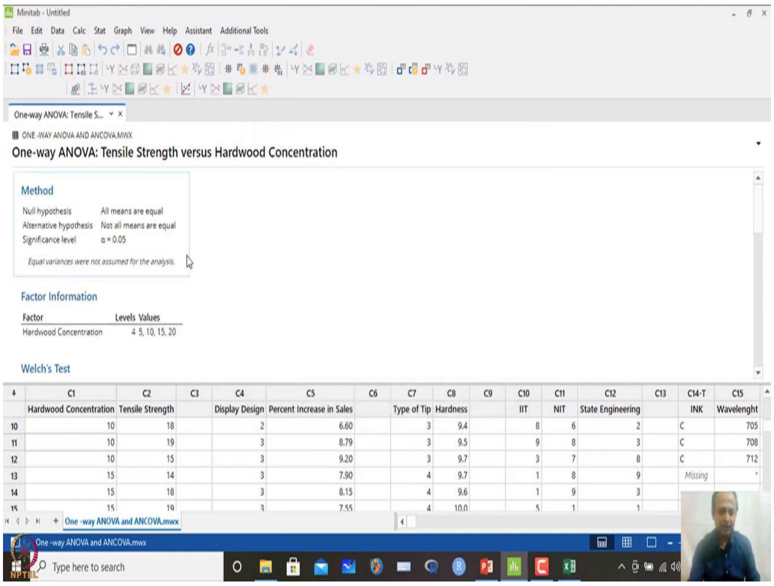


(Refer Slide Time: 21:25)

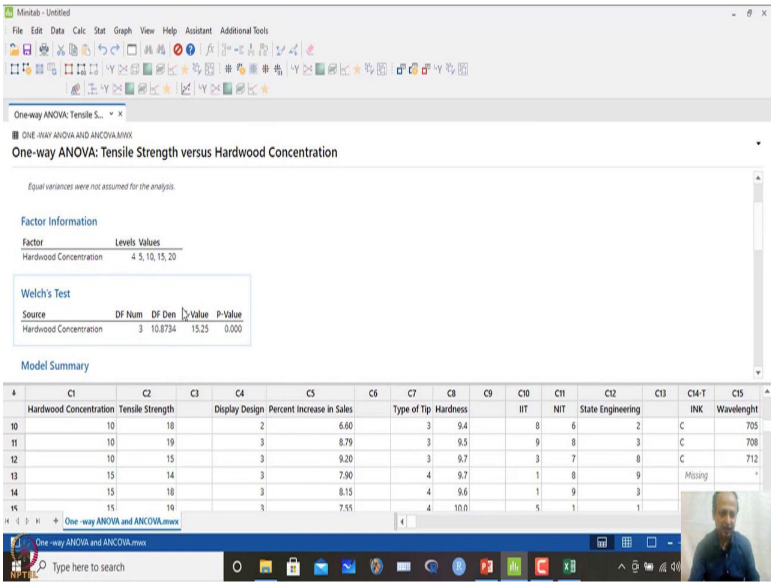


Then in that case what will happen is that another test, Welch’s test, which is equivalent to this one-way analysis will be applicable. So, in the options if you do not assume equal variance or variance is not same, in that case statistical test that exist which is known as Welch’s test.

(Refer Slide Time: 21:33)

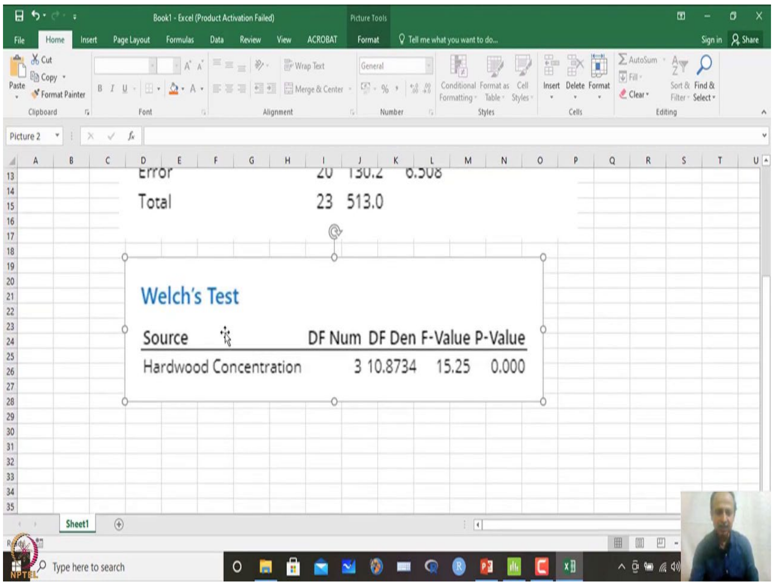


(Refer Slide Time: 21:34)



So, if you click ok over here you will get another value which is given over here and I just copy as picture and I will paste it over here which is equivalent and which is very strong test also which is recommended in case the variance is different. So, this is Welch's test what you see over here.

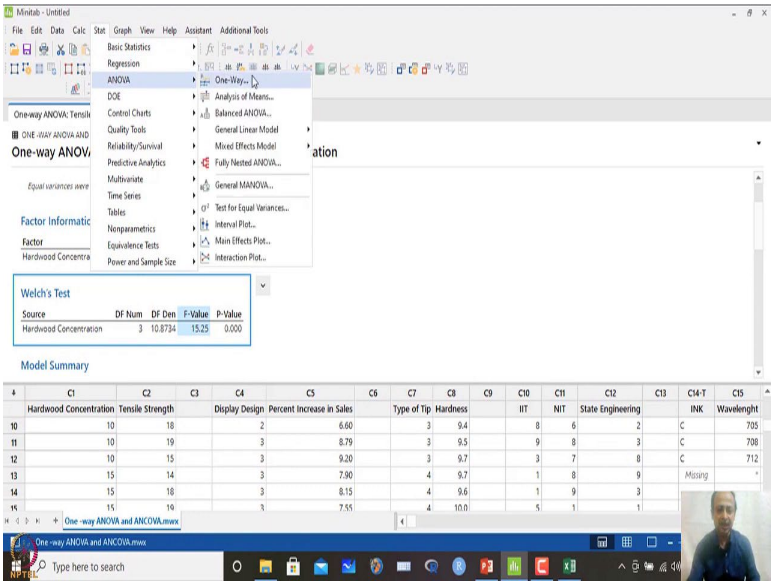
(Refer Slide Time: 21:43)



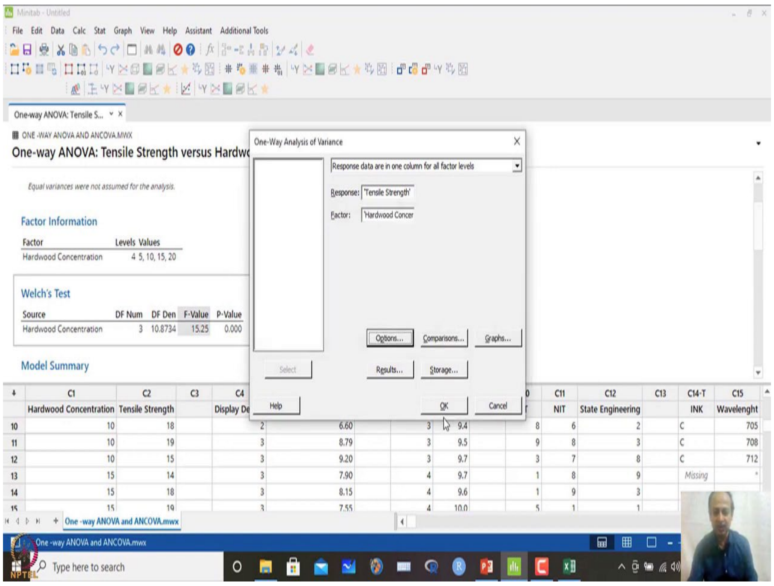
Here also we are checking whether the levels when I am changing whether it is influencing the mean value of CTQs like that and p value is coming out to be less than 0.05. So, that indicates, but this test is only applied when the variance is not same.

So, this is a statistical test which is equivalent like ANOVA analysis like that when the variance at different levels are not same. So, we can use these Welch's test in that scenario here it is not the scenario, but I have shown you the options. So, when you just click over here.

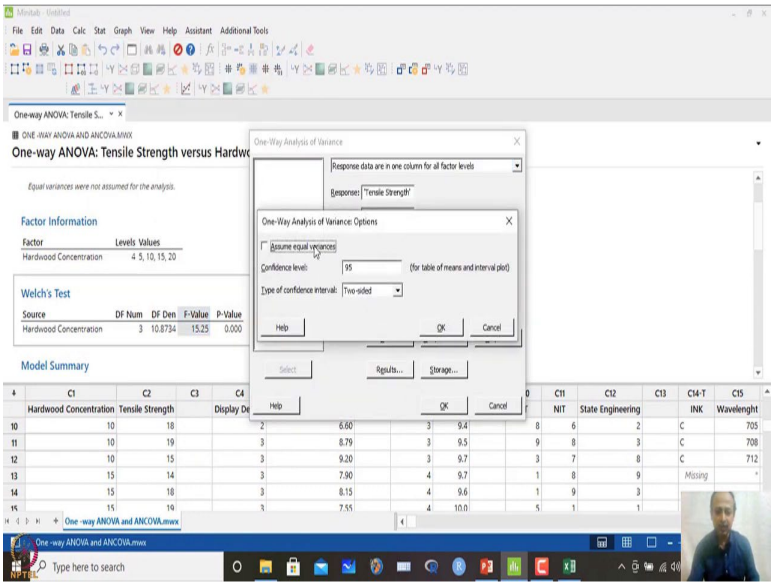
(Refer Slide Time: 22:21)



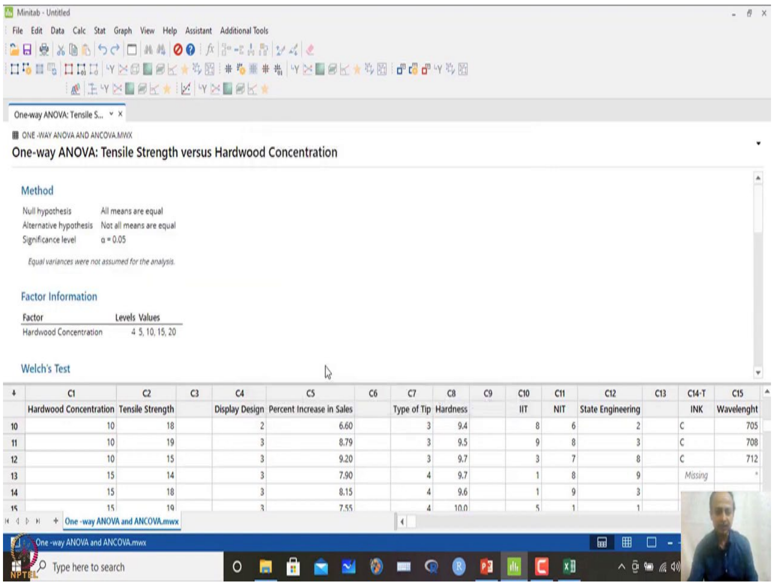
(Refer Slide Time: 22:23)



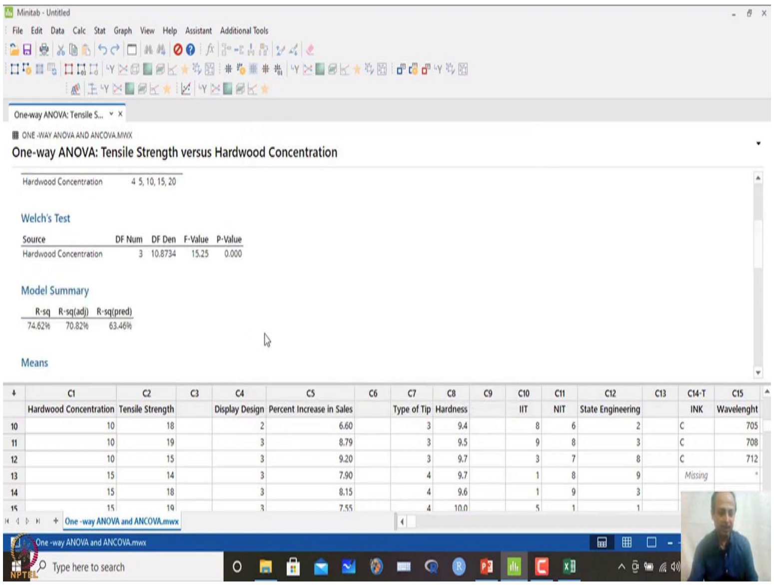
(Refer Slide Time: 22:24)



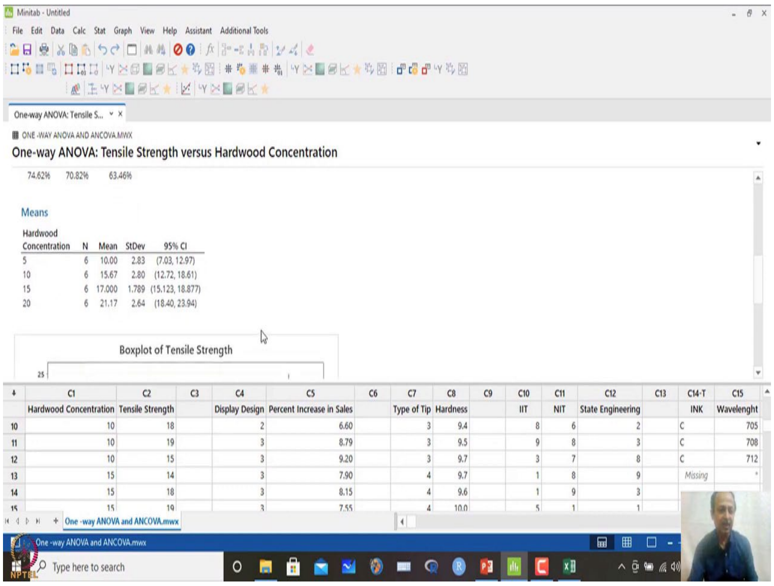
(Refer Slide Time: 22:34)



(Refer Slide Time: 22:35).



(Refer Slide Time: 22:36)



So, if you go to *stat, analysis of variance*, *one-way analysis of variance*, in options if you do not click this one and we tested that variance is not same just uncheck this one. So, if you uncheck this one immediately Welch's test will come. So, the results will be reflected over here.

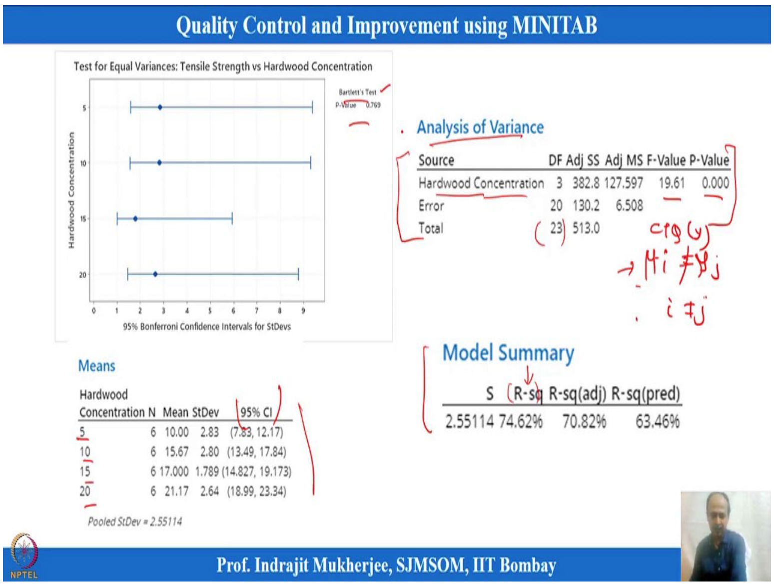
And you will get all other values like that confidence interval and all these things. So, model summaries how much values over here. So, this is at present not required that we are not explaining that one we are explaining that what is the overall idea. So, the at least there are two levels which is which are different basically which are different over here and that is the Welch's test ok. Now, let us go back to another important concept.

So, this what you are seeing over here is basically the analysis of variancetable and this is the same what we have derived using MINITAB and this is the variance test Bartlett test that I have shown over here.

And this is the confidence interval at different levels 5 percent, 10 percent, 15, 20. So, 95 percent confidence interval is given and model summary R^2 values we try to see how much of the variability of total variability is explained basically by hardwood concentration variation when I change the levels. So, that means, whether this factor is very much significant or not that can be seen by this R^2 value. Higher R square value means basically the change in the hardwood concentration is influencing the overall variability basically. So, that is known R^2 value which is known as coefficient of determination that will come when we are discussing about regression analysis.

So, that will be more clear when you see the formulations of regression. So, this we will leave out at present moment model summary is over here. We are interested in this ANOVA analysis, 19.61 and p values over here. This indicates that there is at least one pair of $\mu_i \neq \mu_j$ over here.

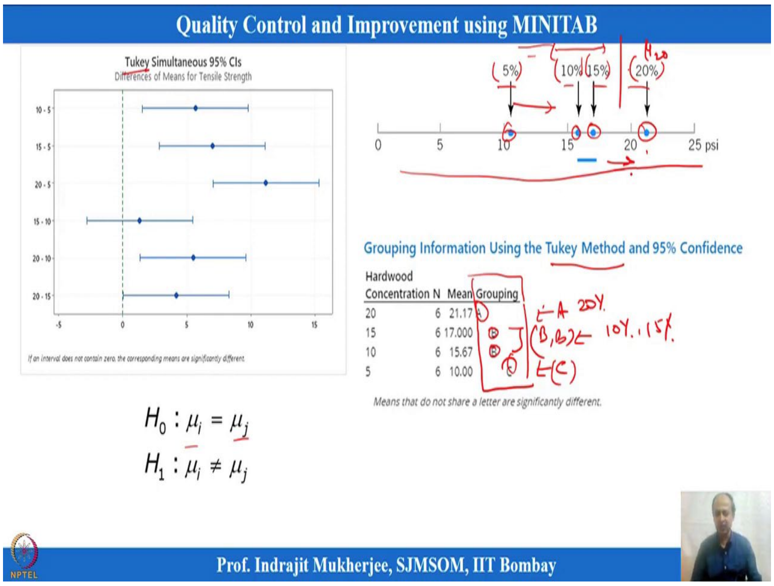
(Refer Slide Time: 24:18)



So, in this case at least there is one i and one j where the mean response is different like that for the CTQ or response basically when I change the level of hardwood concentration.

So, Bartlett test is used if normality assumptions is taken, but in case normality assumptions you are not taking in that case Levene's tests is there, multiple comparison test is there. And the multiple comparison is more powerful than Levene's test. And Levene's test is a non-parametric test which can be also used for interpretation when the data size is small or distribution is skewed. So, in that case we can use Levene's test and later on we will see some scenarios where when it can be applied.

(Refer Slide Time: 25:14)



So, let us try to see that another important concept over here which is shown in this diagram. What you see over here is that I know that two means are different, but which one is different from which one?

Whether it is 5 with 10, 10 with 15 or 15 with 20, that ANOVA analysis cannot tell you and for that we need something which is known as multiple comparison test and there are different methods of doing multiple comparison test.

And we will go by one of the methods which is known as Tukey's method which is given in our MINITAB software and there are other options Fishers method is also given. So, anyway, so, there are different methods, but we will prefer using Tukey's method over here. I will explain one methods or other methods also you can see.

So, over here the overall objective is that which means are similar. Why I am doing this because I want to find out which level I should freeze so, that I get the maximum CTQ and that is the most optimal level.

So, where do I set 5 percent, 10 percent, 15 percent, 20 percent then I need to know which is different from which one. So, over here I need to know whether 20 is different from 15 or 15 is different from 10, 10 is different from 5 like that paired comparison we want to check.

So, in this case I will use in MINITAB. So, now, we have seen that there is significant ANOVA analysis says that there are two levels which are significantly different let us figure out which is different from which one. How do I do that?

So, this will be like paired comparison what I told. So, that will be reflected when I use the Tukey's multiple comparison test, when I am doing that I will get this information. How do I get that? We will see some letter codes that will come over here and we are only interested in seeing the letter codes over here which is written as A, B, C like that.

So, the letter code, which are not similar that levels are significantly different. So, here I am getting a letter code of A and here I am getting a letter code of C and these two levels I am getting a little code of B and B like that. So, when they are same letter code; that means, there is no difference between 10 percent and 15 percent over here.

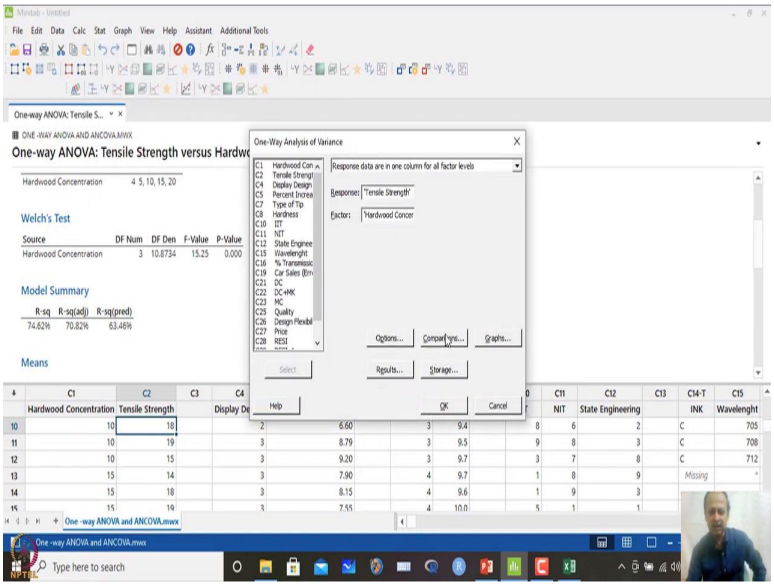
But A level which is 20 percent over here is significantly different from this 10 and 15. Mean of 20 is statistically different from any of the other three basically.

So, C is also far away from this 15 and this is very different from this 15, 10 and 15 like that. So, this is very different. So, at 5 percent we are getting the lowest mean over here and we want to maximize.

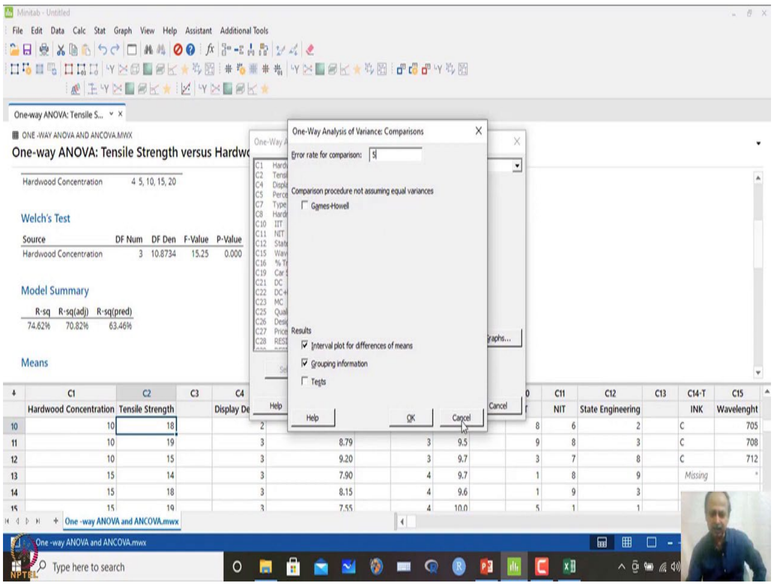
So, it says that which mean is different from which one. So, paired comparison like that this is known as multiple Tukey's multiple comparison test that you can find out when you are doing this in MINITAB, how to do that?

So, what we have to do is that because we have found significant difference here in two levels like that. So, to understand which level is different from which one I go to stat ANOVA analysis again I use one-way analysis of variance.

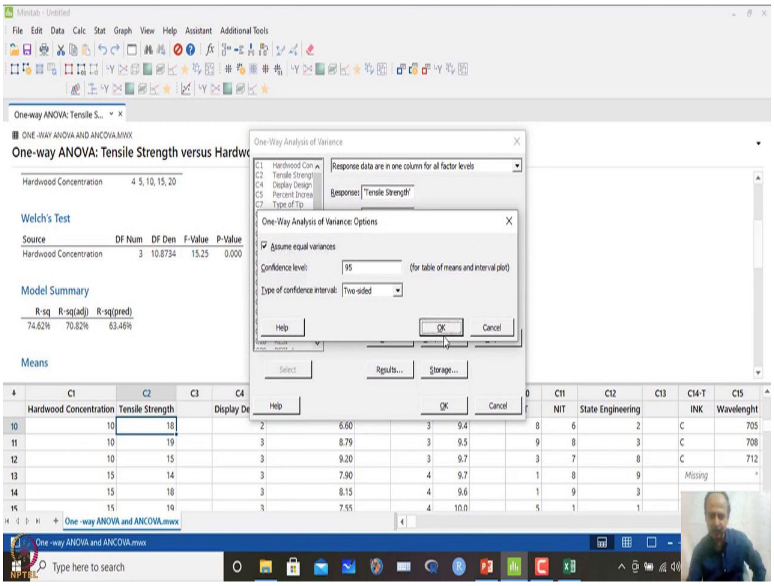
(Refer Slide Time: 28:53)



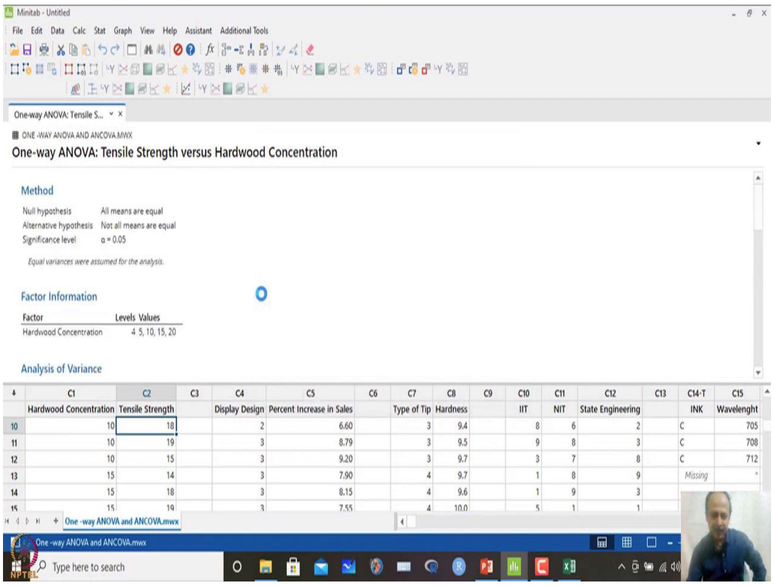
(Refer Slide Time: 28:55)



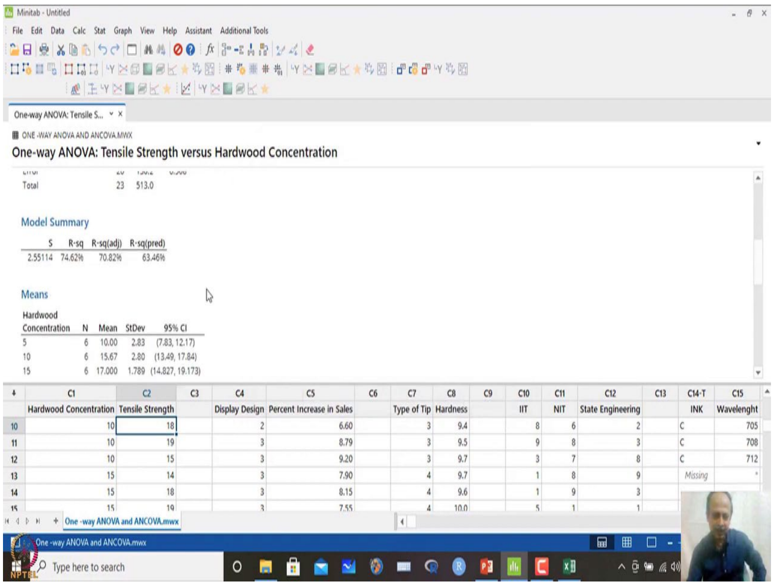
(Refer Slide Time: 29:04)



(Refer Slide Time: 29:11)

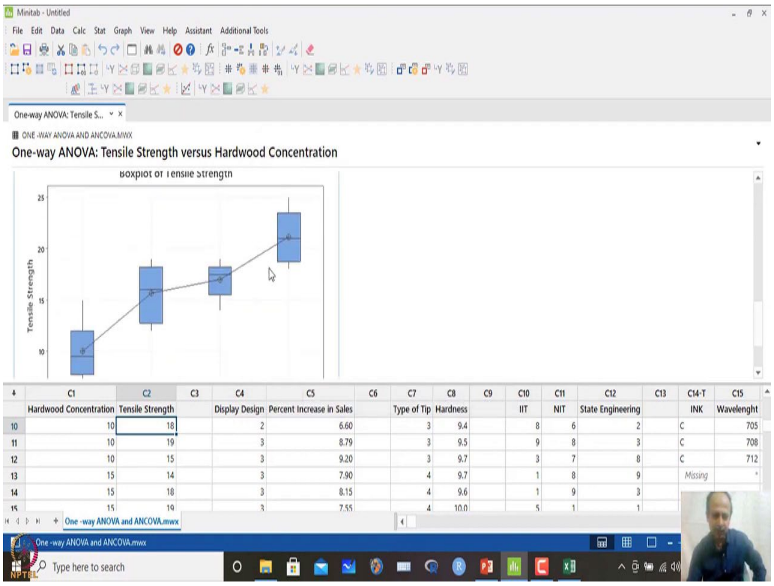


(Refer Slide Time: 29:12)



Now there is a comparison test that is given over here. When you go to comparison test, so, and all this one you keep it as default. So, options over here. So, equal variance if I assume equal variance over here. And then do the test, let us say and we get the analysis over here.

(Refer Slide Time: 29:13)

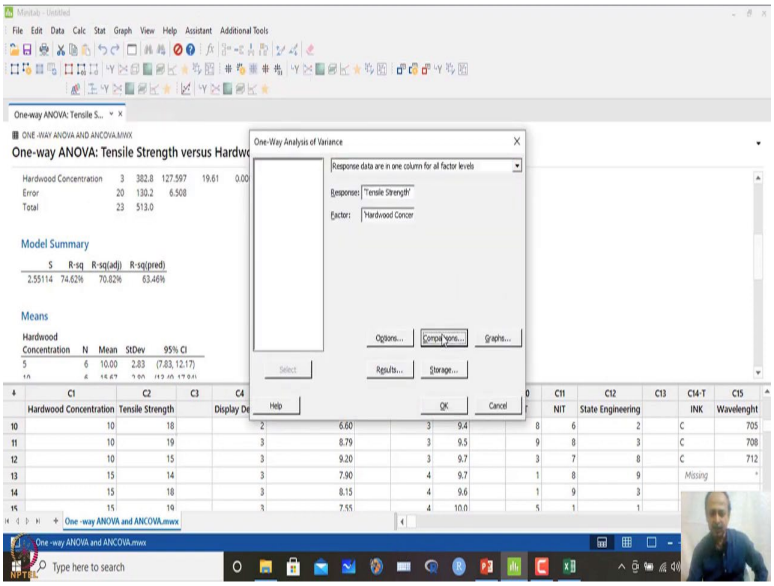


(Refer Slide Time: 29:20)

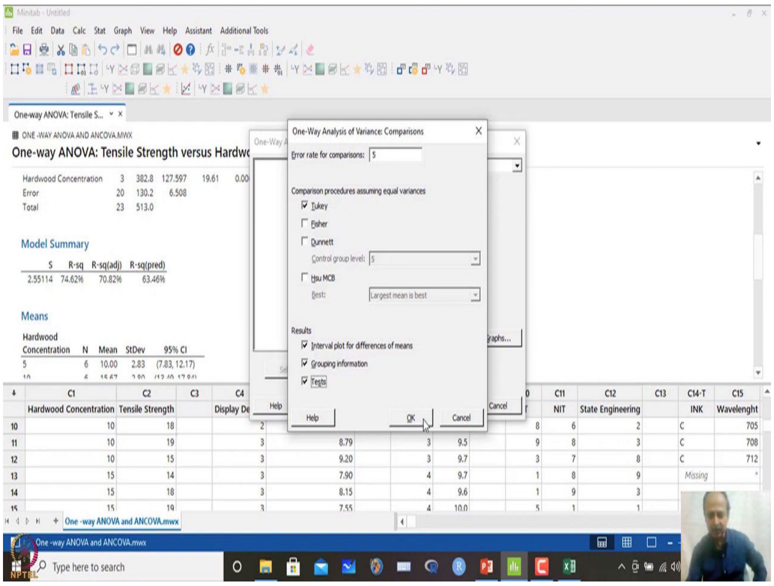
The plot is shown over here you can see the box plot also shown over here. So, this indicates that how the values are changing. So, hardwood concentration is increasing. So, but what you see is that 10 and 15 percent are more or less having overlapping distributions.

So, slope is there when I compare 5 with 10 and 15 and 20 is also having a high slope as compared to 10 and 15 like that and 5 is the lowest one and 20 is the highest one like that.

(Refer Slide Time: 29:53)

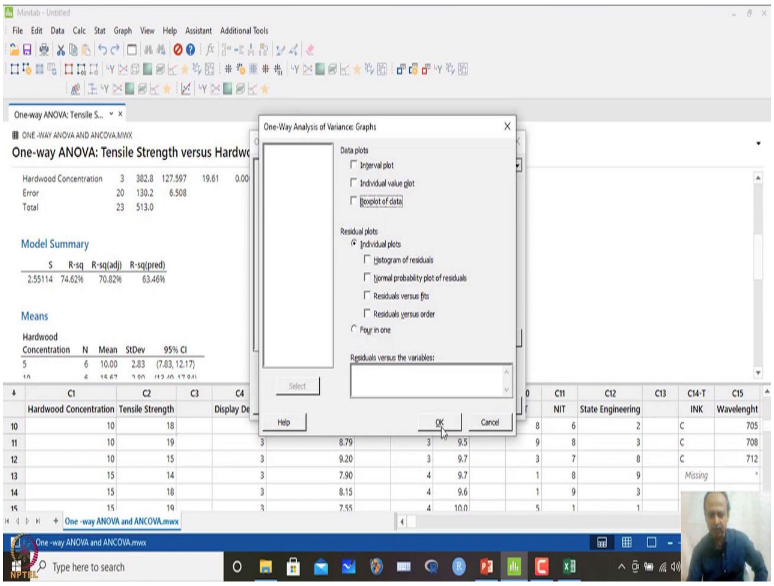


(Refer Slide Time: 29:54)



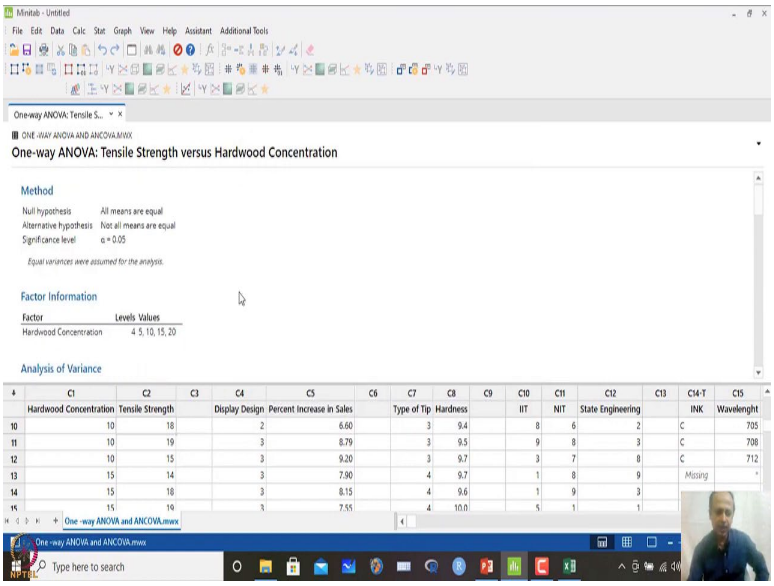
So, box plot will give you that indication over here and then what we have to do is that, I go to *stat*, what we wanted to do is that multiple comparison. So, I go to comparison test over here and I go to Tukey's test over here. I do not see any other test what we will adopt only Tukey's test over here there are other test which can be used, but I am using only Tukey's test. So, grouping information this is very robust test Tukey's test like that.

(Refer Slide Time: 30:15)

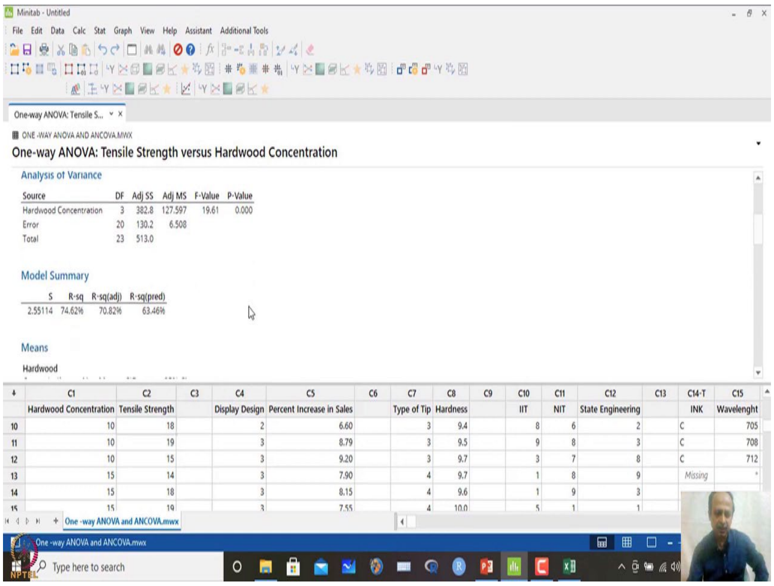


So, grouping information is important for us and then we will click ok and we will not change any other default condition. So, graph we want to check box plot you can see that one otherwise you can ignore that one. If I ignore already box plot we have seen all the data set.

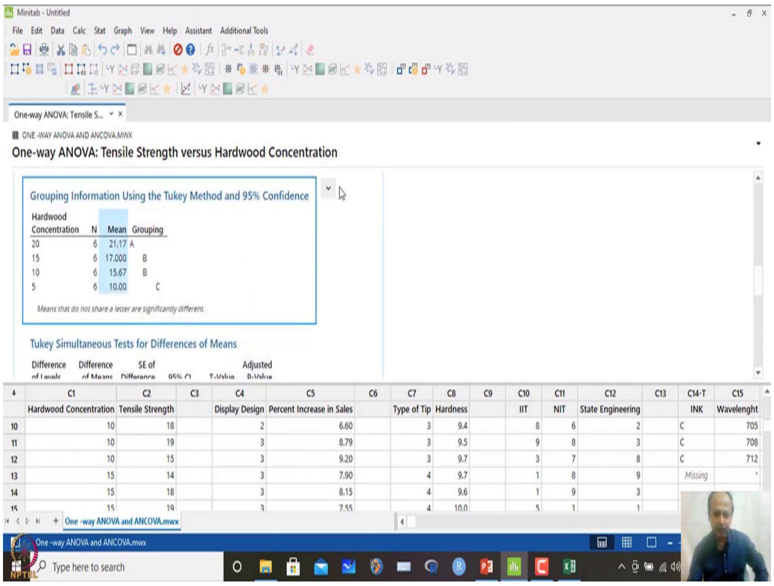
(Refer Slide Time: 30:25)



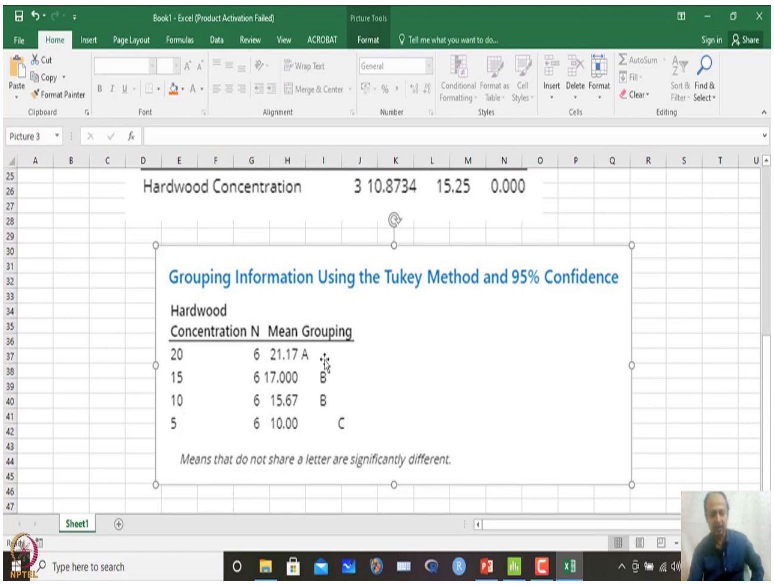
(Refer Slide Time: 30:27)



(Refer Slide Time: 30:30)



(Refer Slide Time: 30:38)



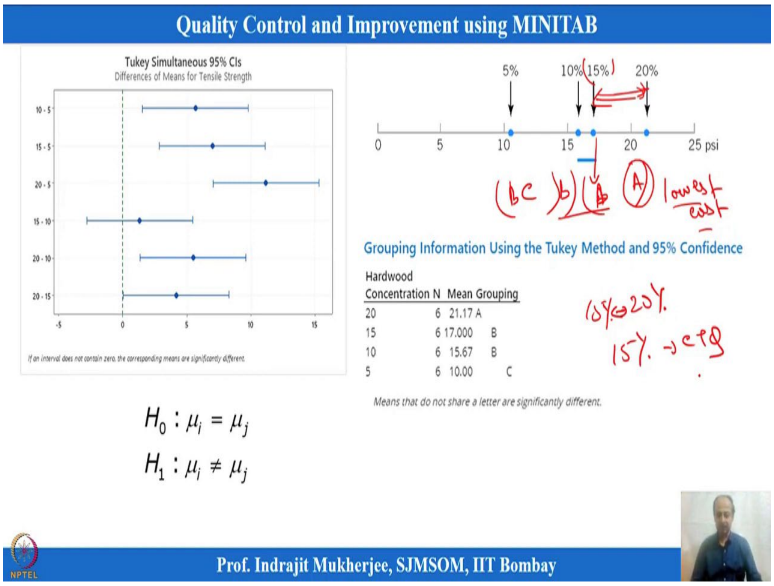
And I click ok what will happen is that, I will get a letter code over here with group information. So, if you click this one go down over here and copy like this and you will get a group information over here. What you see the same results over here, 20 is giving a letter code of A, 15 and 10 is giving a letter code of B and 5 is giving a letter code of C.

So, A is very different from 15, 10 and 5 like that because this letter code is not matching with any of the other levels 5, 10 and 15 like that. So, in this case what we can say is that

A is having a significant higher mean as compared, but B and C is giving me that is 15 and 10 levels over here is giving me the same mean values.

So, if you have to freeze which is the level I will select over here? I will go by the 20 percent hardwood concentration because that is giving me a significant higher tensile strength over here that is given me a significant higher tensile strength over here, but if we have got letter codes.

(Refer Slide Time: 31:39)



That means we have got letter codes over here. So, in case we get letter codes; that means, 20 similar let us say it is giving you a letter code of A and 15 is also giving you a letter code of A and this is B let us assume and this is B like that. So, in this case both are having letter code A; that means there is no statistical difference between 15 and 20 percent.

So, in this case I will go by a lowest cost, I will go by selecting the level which is giving me lowest cost like that because if hardwood concentration 15 maybe this is the lowest cost. And there is no significant difference at population level because hypothesis testing at population level based on the sample information.

So, in this case I will freeze at 15. Here there is statistical difference because this is A, and these others are this one was coming out to be B and this was C like that. So, A is significantly different. So, we should freeze at 20 percent that is the optimal level basically we should freeze assuming this is the only factor and then, but otherwise if both the levels are showing the same letter code, in that case I will go by the lowest cost. So, if 15 and 20 is giving me if I have options selecting within 15 and 20, I will go by 15 percent which will maximize the CTQ values like that and that is the level I will select because 15 and 20 is not statistically different.

So, whether if I freeze at 20 or 15 does not matter only matter what we have to see over here is that which is giving me lowest cost setting like that. So, we will go by the lowest cost setting and overall it is the optimal scenario that is that we are getting over here. So, we will stop over here and we will continue with the assumptions of analysis of one-way analysis of variance.

And further we will discuss some more cases on this before we go into actual more than one factor experimentation that we will discuss in subsequent slides. So, we will stop over here and we will continue from here and try to figure out what are the other things we need to check and do in case while we are doing one-way analysis of variance ok.

So, thank you for listening we will continue the sessions starting with again one-way analysis of variance model adequacy check ok.

Thank you.