

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

**Lecture - 23**  
**ANCOVA and Nonparametric Test**

Hello everyone and welcome to our 23rd session on Quality Control and Improvement with MINITAB, I am professor Indrajit Mukherjee from Shailesh J Mehta School of Management, IIT Bombay. So, in last session we were discussing about analysis of variance and then we discussed that when we are trying to test which factor to be included in the final design of experiment.

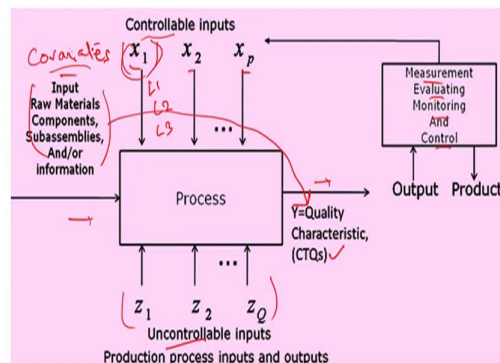
So, for that may be one factor at different levels you are trying to check and analysis of variance is the appropriate technique which we will allow us to understand whether the when I change the factor whether the mean value of Y is significantly changing or not. So, we have taken some examples also and we have also seen that how to check the model adequacies like that.

So, if error follows normal distribution or there is a heteroscedastic scenarios existing or not or, and that situation what is to be done that also we have discussed. Durbin Watson statistics also we have discussed which can be used for testing auto correlation between the errors or errors are independent or not.

So, today's lecture we will concentrate on extended version of this. So, analysis of variance in certain situation what happens is that you may have a continuous variable. If I have a discrete factor X and I want to see if I change the X level what is the influence on Y. But, there can be variables like temperature which I cannot control and it can take some values. So, it does not have discrete values, but it can have influence on the output of the process.

(Refer Slide Time: 01:54)

### Quality Control and Improvement using MINITAB



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



So, let me just show you what I mean to say. So, this is the overall diagram process diagram that we have concentrated. There will be control variables, uncontrollable or uneconomical factors which also changes and then this is the CTQ which is coming out of the process basically.

So, one component enters into this process and goes out of this, we measure the CTQs and try to check whether everything is going fine and monitoring control that one. So, these are the settings which we change  $x_1, x_2$  up to  $x_p$  and this is uncontrollable variables or difficult to control.

So, in this case and there can be also inputs that keep on changing that means it is coming from the earlier process. So, this variables which I cannot control, but these has influence on my characteristics Y over here, this can some of the terms that is used is known as covariates this is also known as covariates ok.

So, this variables can take different values and we do not have any control, but we know that this influences the final CTQ or Y and we want to control X at different levels. And I am basically interested not on this covariates I am interested whether the X factor influences on Y or not. Later on I can take care of that covariates by different means in design of experiments, but what we want to know is that whether X influences Y or not.

(Refer Slide Time: 03:20)

**ANCOVA**  
**(Analysis of Covariance)**  
**GENERAL LINEAR MODEL**



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



So, for that there is another analysis which is an extension of analysis of variance which is known as ANCOVA which is Analysis of Covariance and for this, what is required is that we use a General Linear Model over here. For the general linear model here regression which is another important concept which we will discuss just next after this and which is extensively used in the design of experiments to understand the relationship between Y and X and also to screen factors. For e.g. which factor to be selected in case X is continuous. So in that case what we do is that we use regression technique. Here we are talking about linear regression models and we will not go into discussion of linear regressions model after this.

But how to use the ANCOVA results, how to interpret that results we are only interested to understand that there is a continuous variable there is a factor I want to understand whether the factor influences my outcome that is a CTQ or not. So, one of the examples we will take over here and so that it is quite easy for you to understand.

So, what I am saying is that this covariate influences my final outcome, but I am testing one factor at different level. So, this can be having level 1, level 2, level 3 which are discrete levels. But these covariates along with whenever I am changing covariates values are also changing and I want to see that in presence of covariates whether X influence I can detect on Y.

(Refer Slide Time: 05:07)

## Quality Control and Improvement using MINITAB

### NONPARAMETRIC ANOVA

#### Kruskal-Wallis Test

(Data for all of the groups have similarly shaped distributions)

#### Mood's Median Test

(Data for all of the groups does not have similarly shaped distributions or there is influence of Outliers)



Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



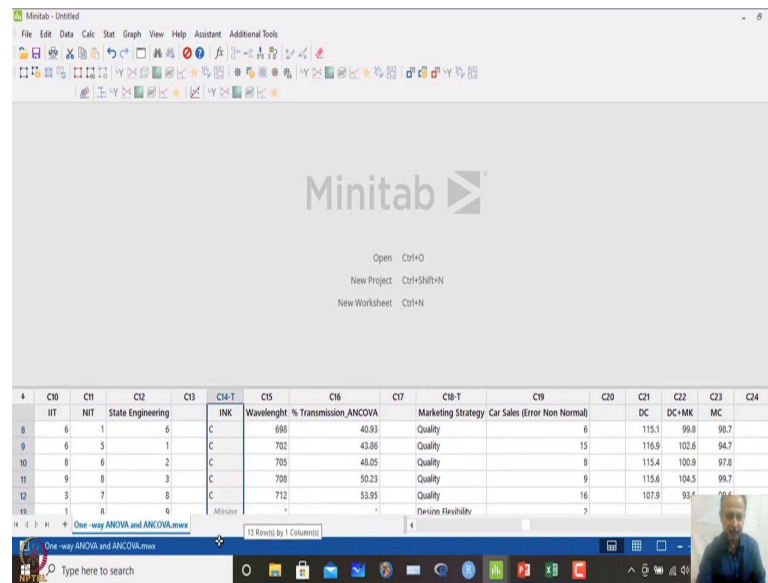
So, in this case what we will do is that? We will take some examples and one of the examples we will take that is already we have with us.

(Refer Slide Time: 05:19)

	C10	C11	C12	C13	C14-T	C15	C16	C17	C18-T	C19	C20	C21	C22	C23	C24
1	3	1	3	A	690	47.16	Quality	3	118.8	105.4	102.1				
2	1	2	4	A	705	52.83	Quality	6	122.6	101.1	105.8				
3	1	1	4	A	712	58.63	Quality	7	115.6	102.7	99.6				
4	3	2	6	B	695	38.47	Quality	4	113.6	97.1	102.7				
5	4	3	3	B	702	45.36	Quality	6	119.3	101.9	101.9				
6	7	4	5	B	708	57.56	Quality	7	115.9	98.0	101.9				

So, I will go to the file where this example exist. So, here what we see is that C14 column C15 and C16 gives you the information about different types of ink and percentage transmission which is basically Y over here. So, this is a percentage transmission that is Y over here, this is the CTQ which I have monitoring over here. And I want to see the influence of ink A B C categories over here.

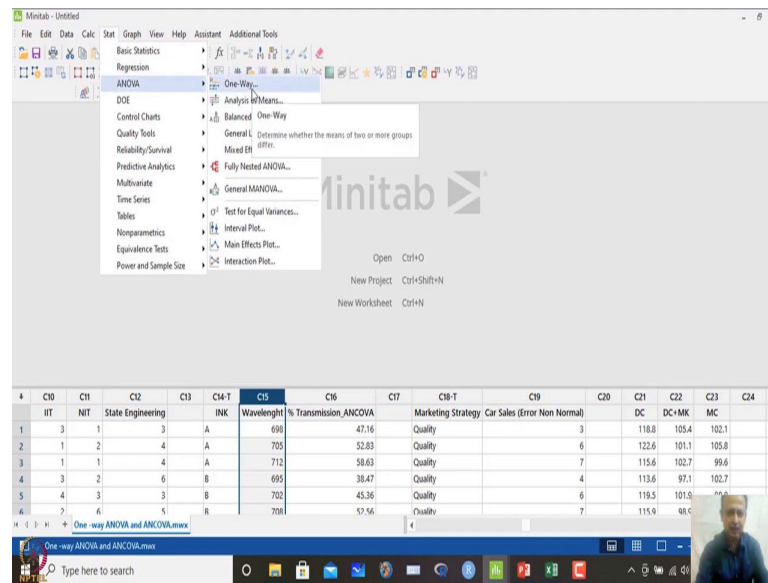
(Refer Slide Time: 05:44)



So, in these case, whether the ink is changing the percentage transmission or expected value of percentage transmission over here? ink is discrete variables at different levels discrete levels, we are checking and I want to see whether it influences percentage transmission over here.

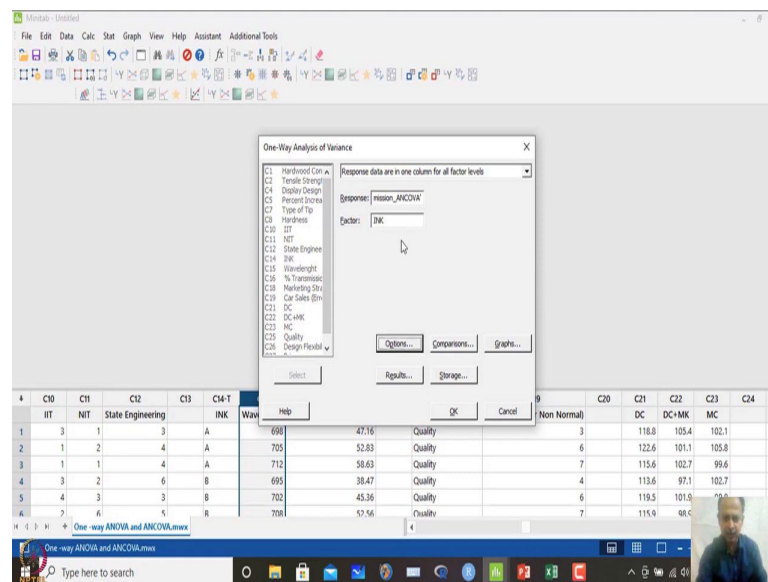
There is another important factor which is a covariates over here which is also changing accordingly when I am changing. We cannot control this one, but this value keeps on changing when I change the levels of ink. Intuition says that this can influence my percentage transmission from my experience or previous knowledge and also maybe theoretically we also understand that this influences. But at present, as we cannot control this, we can treat it as uncontrollable one. So, in this case, but whether the factor is important or not that is basically what we want to see ok.

(Refer Slide Time: 06:48)



Say if you see C14 and C16 if I ignore this one C15 and I do one way analysis of variance what is the results we can see. So, I am doing a one way analysis of variance.

(Refer Slide Time: 06:54)



And in this case percentage transmission is the response and the factor over here what we have consider is ink over here. I want to see that whether analysis of variance is significant or not. I will just check this one.

(Refer Slide Time: 07:14)

**Method**

Null hypothesis: All means are equal  
 Alternative hypothesis: Not all means are equal  
 Significance level:  $\alpha = 0.05$   
 Rows utilized: 1

Equal variances were assumed for the analysis.

**Factor Information**

Factor	Levels	Values
INK	3	A, B, C

#	C10	C11	C12	C13	C14-T	C15	C16	C17	C18-T	C19	C20	C21	C22	C23	C24
	ITT	NIT	State Engineering		INK	Wavelength	% Transmission_ANCOVA		Marketing Strategy	Car Sales (Error Non Normal)		DC	DC+MK	MC	
1	3	1		3	A	698	47.16		Quality		3	118.8	105.4	102.1	
2	1	2		4	A	705	52.83		Quality		6	122.6	101.1	105.8	
3	1	1		4	A	712	58.63		Quality		7	115.6	102.7	99.6	
4	3	2		6	B	695	38.47		Quality		4	113.6	97.1	102.7	
5	4	3		3	B	702	45.36		Quality		6	119.5	101.9	102.7	
6	7	6		1	R	708	57.56		Quality		7	115.9	101.9	102.7	

(Refer Slide Time: 07:16)

**One-way ANOVA: % Transmission\_ANCOVA versus INK**

Source	DF	Adj SS	Adj MS	F-Value	P-Value
INK	2	60.22	30.11	0.75	0.500
Error	9	361.32	40.15		
Total	11	421.54			

**Model Summary**

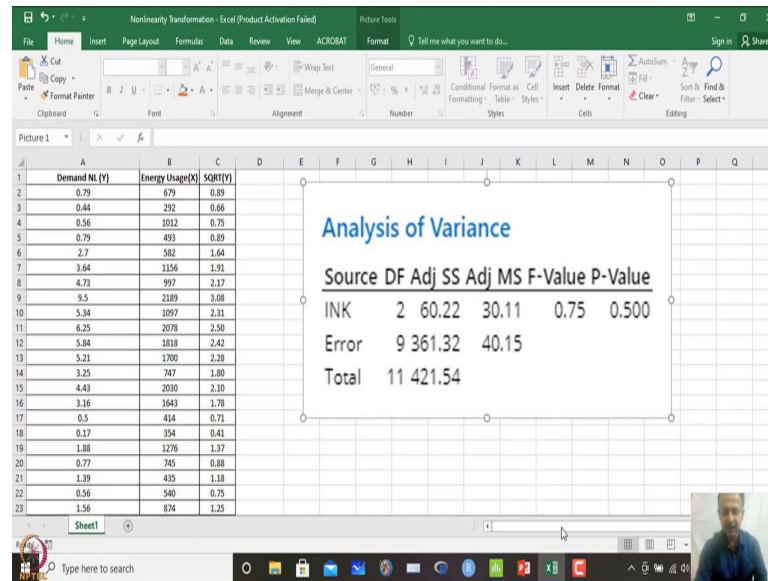
S	R-sq	R-sq(Adj)	R-sq(Pred)
6.33814	14.29%	0.00%	0.00%

**Means**

INK	N	Mean	StDev	95% CI
A	3	52.87	5.74	(44.60, 61.15)
B	3	48.75	10.50	(27.84, 69.66)
R	1	57.56	0.00	(57.56, 57.56)

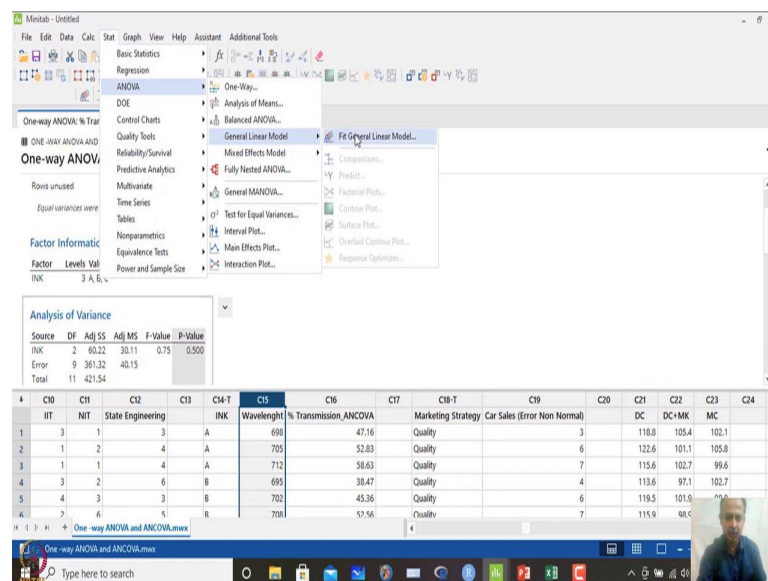
#	C10	C11	C12	C13	C14-T	C15	C16	C17	C18-T	C19	C20	C21	C22	C23	C24
	ITT	NIT	State Engineering		INK	Wavelength	% Transmission_ANCOVA		Marketing Strategy	Car Sales (Error Non Normal)		DC	DC+MK	MC	
1	3	1		3	A	698	47.16		Quality		3	118.8	105.4	102.1	
2	1	2		4	A	705	52.83		Quality		6	122.6	101.1	105.8	
3	1	1		4	A	712	58.63		Quality		7	115.6	102.7	99.6	
4	3	2		6	B	695	38.47		Quality		4	113.6	97.1	102.7	
5	4	3		3	B	702	45.36		Quality		6	119.5	101.9	102.7	
6	7	6		1	R	708	57.56		Quality		7	115.9	101.9	102.7	

(Refer Slide Time: 07:26)



And then I will see the analysis of variance table. What I see over here is that p value is 0.5. So, what we are seeing is that ink is not a factor that is influencing my percentage transmission over here. So, ink is not a significant factor if I am not considering that covariates over here. We are not considering covariates we are doing one way analysis of variance and p value is not coming out to be significant. Now, if I consider this as a covariate and do the analysis what will happen that we have to see. So, for this one way analysis of variance does not have any options. So, we cannot do this analysis over here.

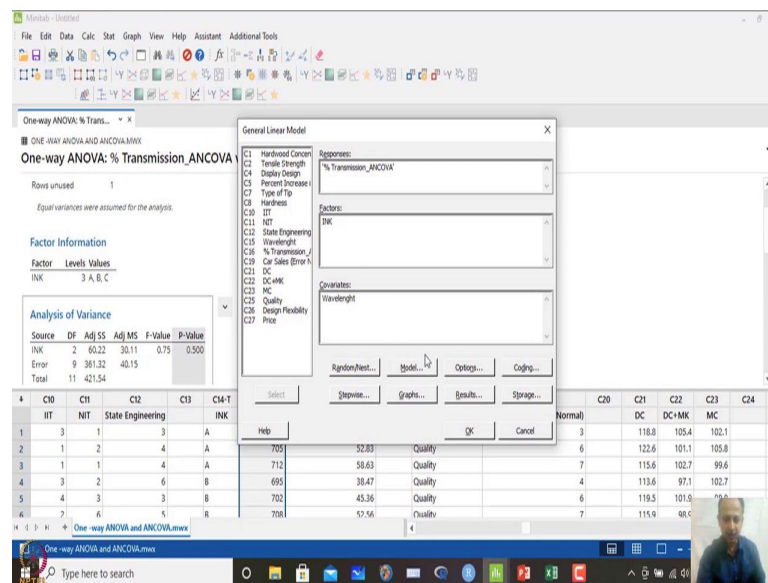
(Refer Slide Time: 08:11)





What we have to do is that we have to go to here and there is a general linear model that is available over here. So, general linear model what we will do is it that fit general linear model over here.

(Refer Slide Time: 08:18)

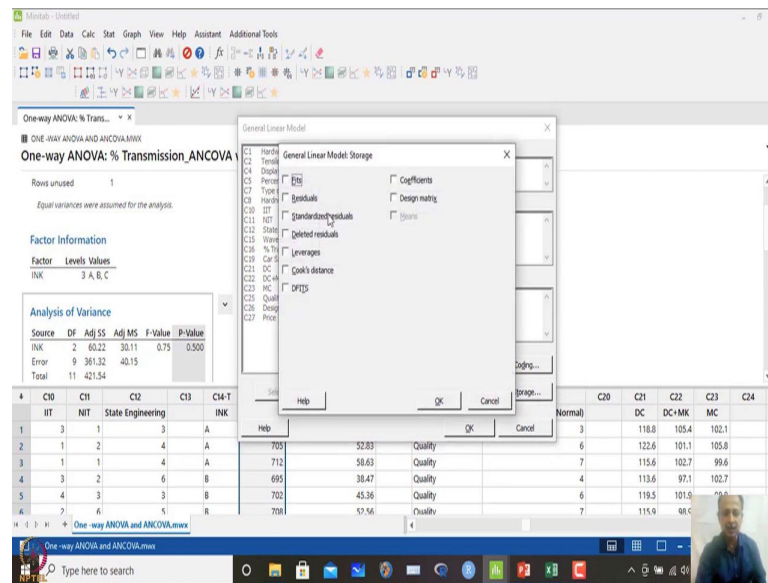


Then, what we will do is that using regression as a underline theme to develop this ANOVA analysis over here. So, anyhow we will discuss about regression after just after this one and that will more clear how regression is used to develop the models like that.

Let us assume that we want to see only the analysis of variance and we are interpreting that from the p value interpretation, I want to see whether the factor influences or not. For this there is a option over here you see covariates one option is given over here in general linear model. So, I can include factor, I can include covariates, I can include response.

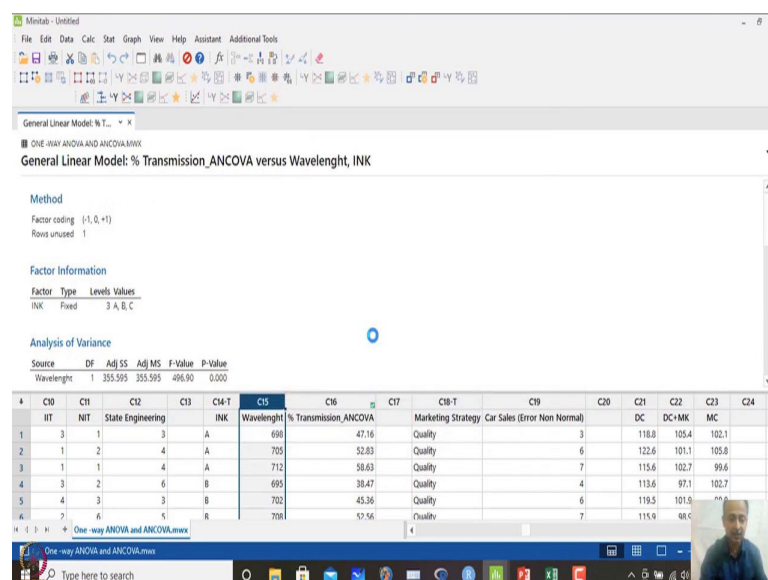
So, what will be my response, percentage transmission is my response, factor will be ink over here which I going to understand and the covariates that we are considering over here is the wavelength, which is the covariates over here there are many more options over here.

(Refer Slide Time: 09:04)

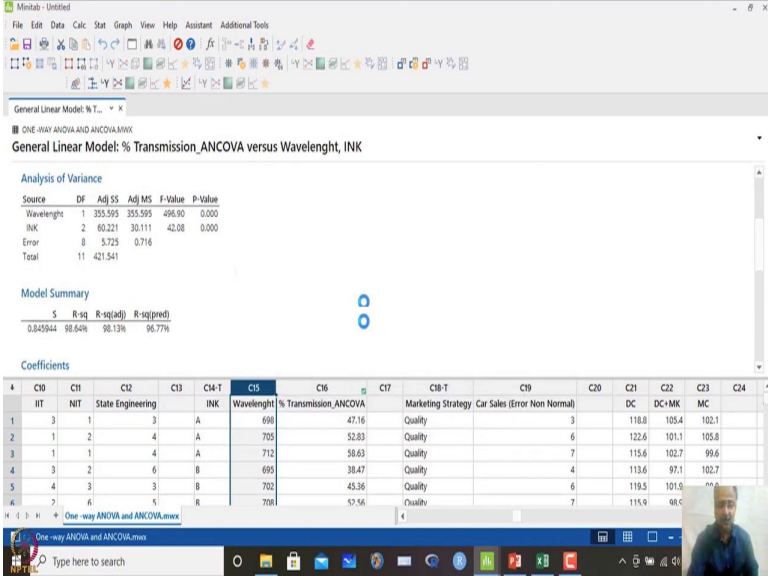


We can store the residual and we can analyze model adequacy we can do that ok. So, nothing else is required over here at this stage because this is using linear regression and we do not understand regression at present. But we want to use and see that whether this covariates I know covariate this variable wave length has an influence. So, but I want to understand only whether the factor is important or not for screening like that.

(Refer Slide Time: 09:34)



(Refer Slide Time: 09:35)



General Linear Model: % Transmission\_ANCOVA versus Wavelength, INK

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Wavelength	1	355.595	355.595	496.90	0.000
INK	2	60.221	30.111	42.08	0.000
Error	8	5.725	0.716		
Total	11	421.541			

Model Summary

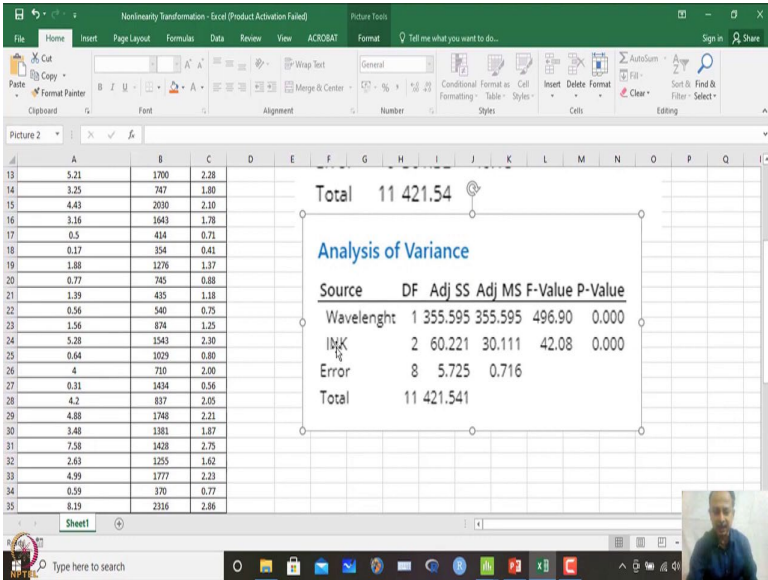
S	R-sq	R-sq(Adj)	R-sq(Pred)
0.845944	98.64%	98.13%	98.77%

Coefficients

IT	C10	C11	C12	C13	C14-T	C15	C16	C17	C18-T	C19	C20	C21	C22	C23	C24
INT	INT	State Engineering			INK	Wavelength	% Transmission_ANCOVA		Marketing Strategy	Car Sales (Error Non Normal)		DC	DC+MK	MC	
1	3	1		3	A	698	47.16		Quality		3	118.8	105.4	102.1	
2	1	2		4	A	705	52.83		Quality		6	122.6	101.1	105.8	
3	1	1		4	A	712	58.63		Quality		7	115.6	102.7	99.6	
4	3	2		6	B	695	38.47		Quality		4	113.6	97.1	102.7	
5	4	3		3	B	702	45.36		Quality		6	119.5	101.9	102.7	
6	7	6		1	R	708	57.56		Quality		7	115.9	100.0	102.7	

So, what I will do is that I will click ok over here and I want to see the analysis of variance table only. So, whenever I have done there is a analysis of variance table, which you can see I am just copying this and I want to paste this one and see what happens ok.

(Refer Slide Time: 09:47)



Nonlinearity Transformation - Excel (Product Activation Failed)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Wavelength	1	355.595	355.595	496.90	0.000
INK	2	60.221	30.111	42.08	0.000
Error	8	5.725	0.716		
Total	11	421.541			

So, parallelly we have earlier done this one that this is 0.5. But when I have incorporated this wavelength what you see that ink factor over here is having a p value which is less than 0.05. Which was not significant when I had analyzed ink and percentage transmission.

So, in this case what we are seeing is that if wavelength and ink values are changing, they are influencing my outcome that is CTQ. So, degree of freedom is 3 minus 1 that is 2. And this is treated as a regression variables over here, wavelength over here and so, in this case what it says is that when I consider this covariates ink is coming out to be very prominent factor over here. So, whenever I have a covariate information about covariates I should include that one to understand whether the actual factor, which I am changing, is influencing Y or not.

So, in this case what we are seeing is that if I ignore the covariates what will happen is that ink is not coming out to be prominent factor. But when I include the covariates over here, so some part of variability if I can segregate the variability total variability into ink and wavelength over here, what I am seeing is that ink is coming out to be prominent over here.

So, that is the way we should interpret analysis of covariance. This one is important thing that I wanted to say. We have to also understand that normality assumptions are model adequacy may not work in certain scenarios, although the analysis of variance is very robust and researchers claim that it is very robust.

But, we have to also understand that real life scenario is not this and there can be deviations, but analysis of variance is very strong and small deviations or moderate deviation does not influence the results as such and my conclusion will be a more or less correct in that.

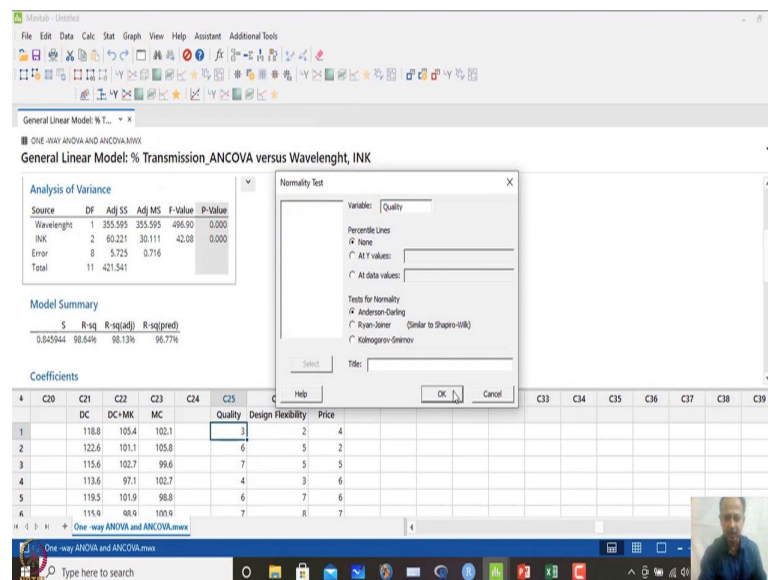
But there are options like nonparametric options that we can adopt in case assumptions are not valid or assumptions we are not able to satisfy or transformation is not working like that. So, in that case when everything fails I cannot adopt the assumptions like normality assumption, heteroscedasticity, many scenario it can happen.

So, in that case there is the alternative test which is know and Kruskal-Wallis Test. So, just an alternative to analysis of variance what we have is Kruskal Wallis test which can be used when groups have similar distribution. Groups means a category if I have three levels A B C like that. So, every category the Y characteristic is following a same distribution.

So, in that case we can use that one and we can check that one. So, and we have done that also in previous cases ok. So, and the other if the shapes of the distribution are different in that case we can also go for Mood's Median Test which is also a nonparametric test.

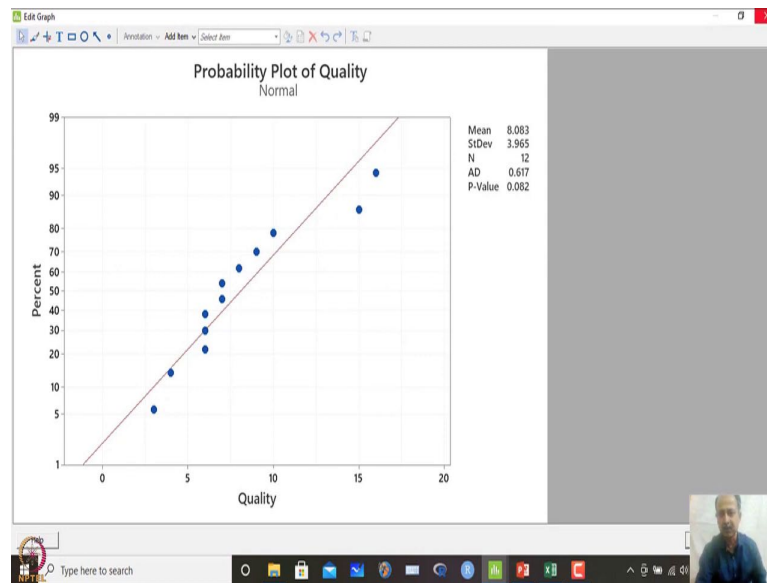
So, if it is similar distribution for the groups I can use Kruskal Wallis test and if it is not that we can use a Mood's Median test over here. So, in this examples what we have done is that we have taken all here also marketing strategy car sales and it was found to be non normal we can do transformations over here. And on transform why we have seen whether the factor is influencing or not based on that we have made a judgment like that.

(Refer Slide Time: 13:26)



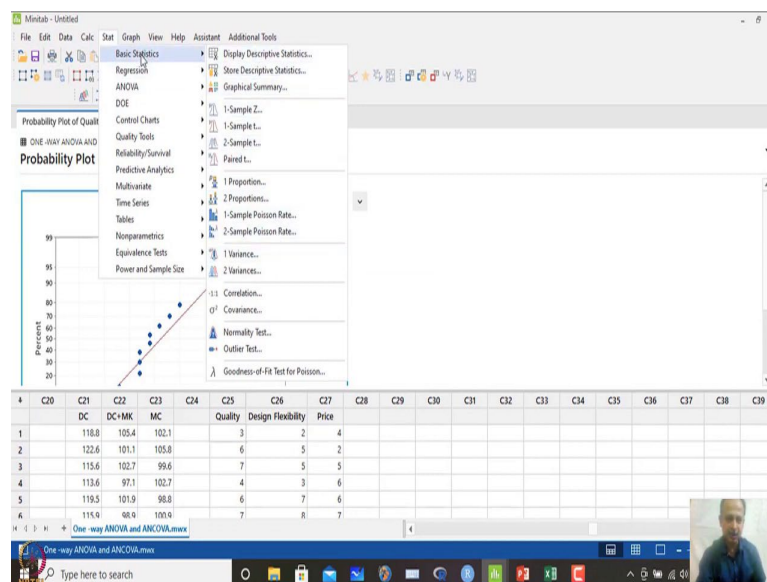
So, let us assume I want to go for non parametric test. What we have seen is that this is quality design and price over here these are the three variables and these are the sales information that we have. And if you can see test this one whether this is normal distributed each of the groups if you if they are normal distributed or not. So, in this case quality we can take Anderson-Darling test is used and what we are seeing is that.

(Refer Slide Time: 13:44)

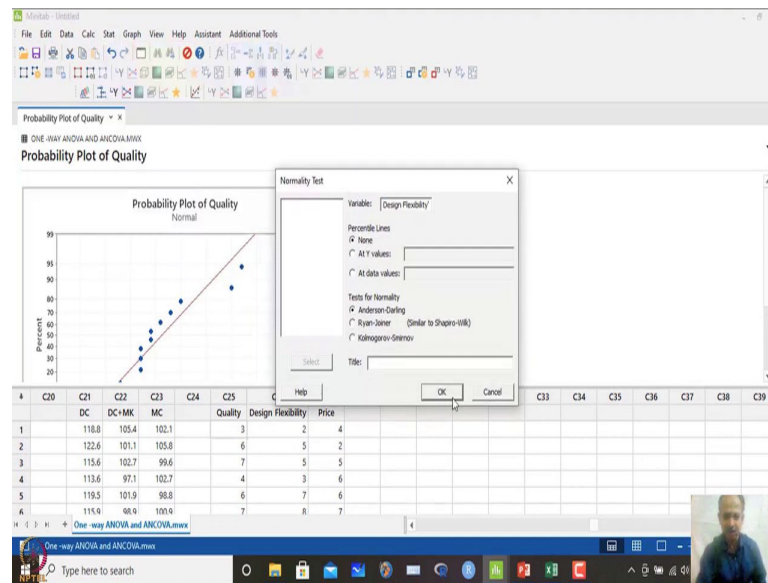


Now, we see that it is 0.08 approximately and that is we can assume that normal distribution assumption is fulfilled over here. Similarly, what we can do is that we can see the next one variables and this we have already shown.

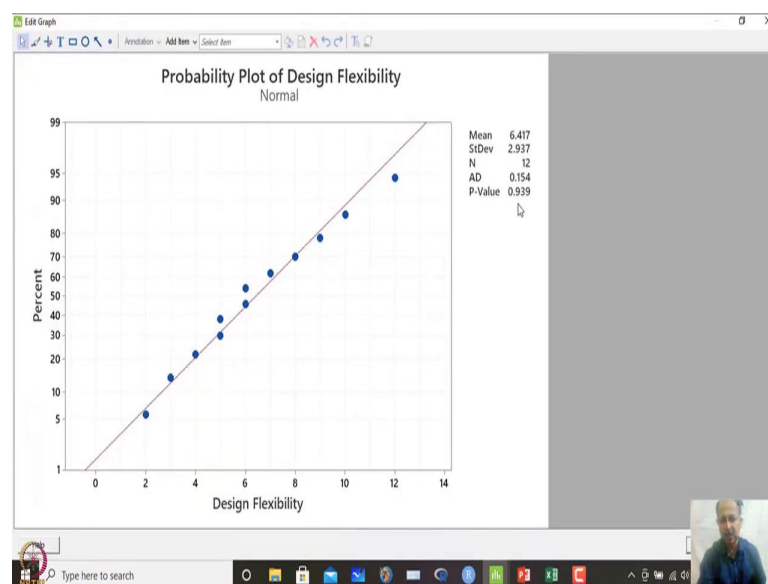
(Refer Slide Time: 13:58)



(Refer Slide Time: 14:04)



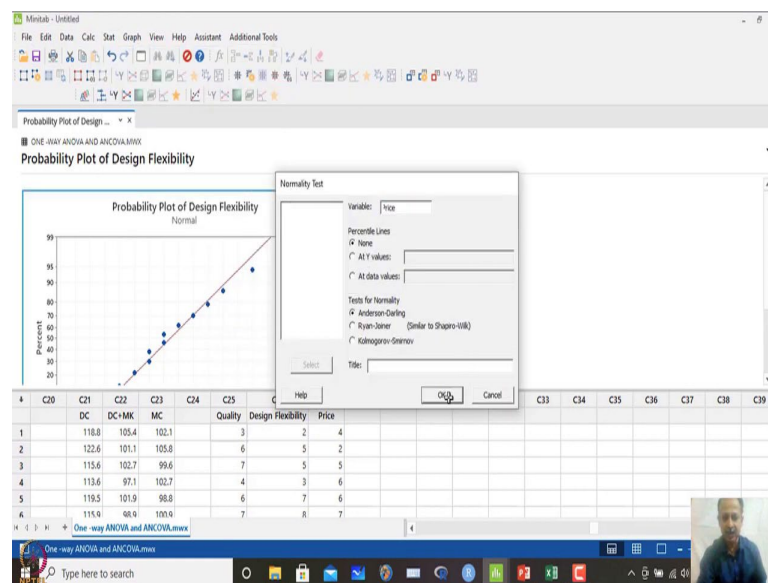
(Refer Slide Time: 14:05)



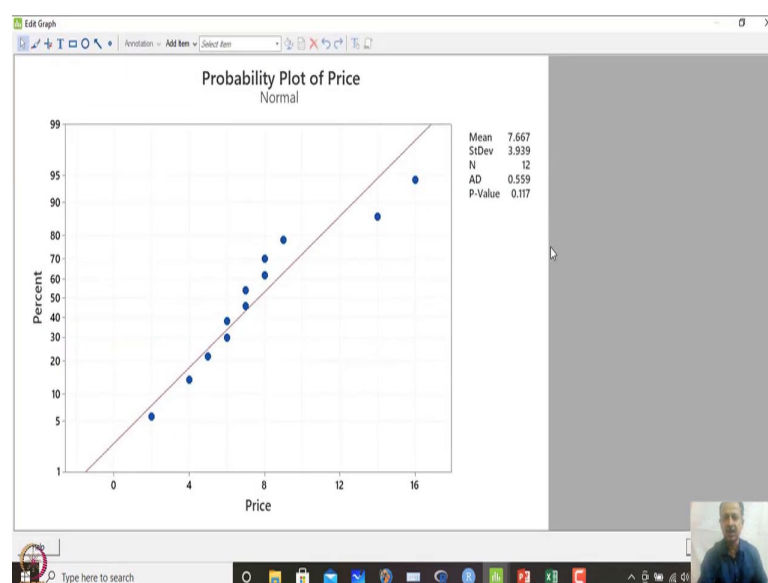
So, we can just see the next one and we can do the normality test for the next variables like that. So, that is design flexibility and I go ahead and I see that the p value that we are getting over here is 0.939 that is also satisfactory.



(Refer Slide Time: 14:24)



(Refer Slide Time: 14:26)

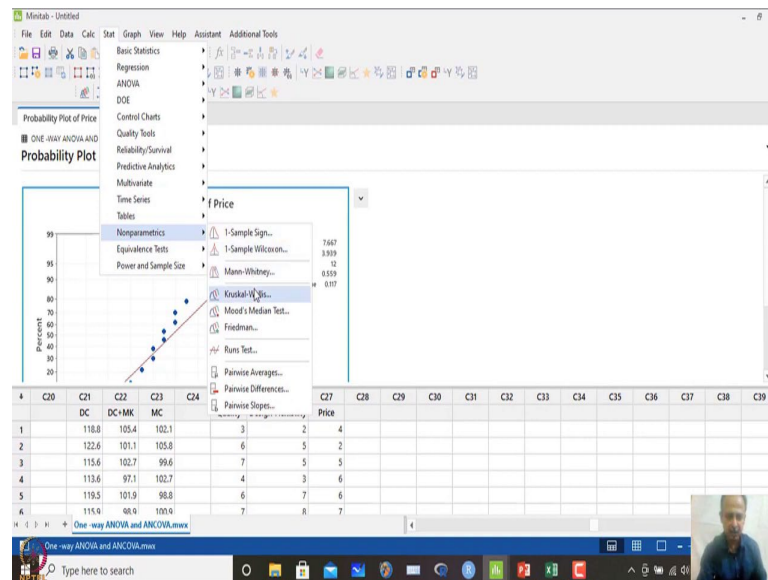


So, for price, Anderson-Darling test is done and here also what we are seeing is that P value is 0.117 over here. So, here also it is satisfied. So, within groups more or less the distribution is more or less the same.

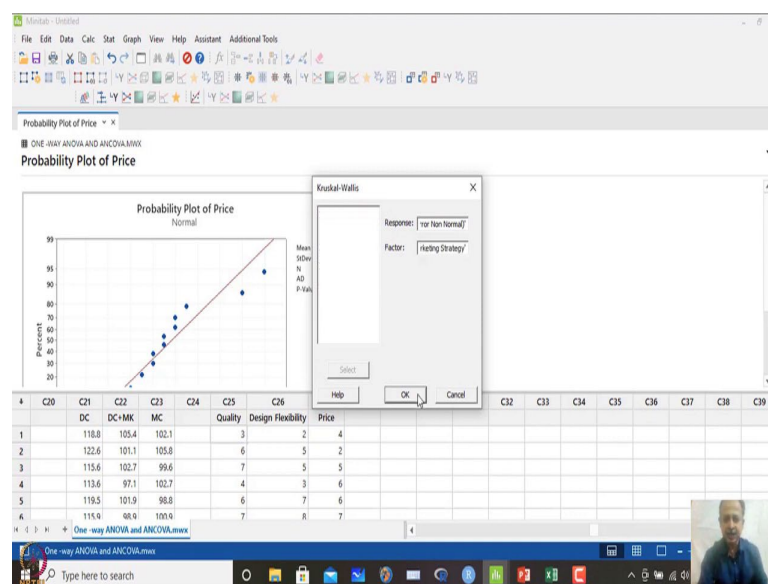
And this can be verified whether it is normal or other distributions also we can verify and that is that option is available in MINITAB also to check which distribution it follows. So, anyhow, so to keep it simple what I am doing is that assuming this one and assuming that transformation and everything fails.



(Refer Slide Time: 14:51)

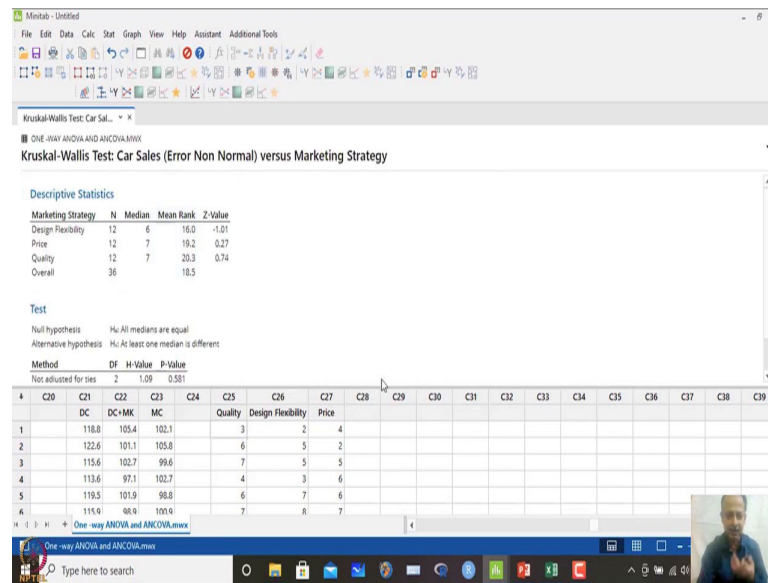


(Refer Slide Time: 14:53)

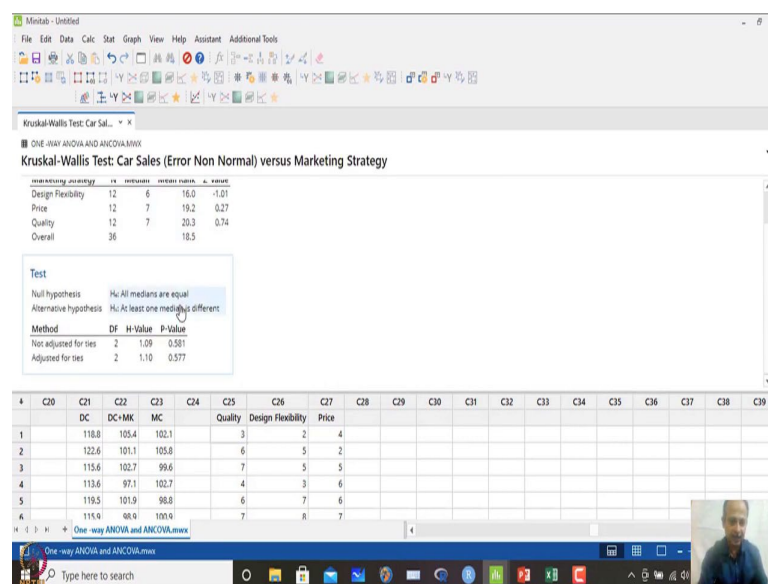


So, I can go to non parametric test and I can go to Kruskal Wallis test over here and what I will do is that I will give the car sales as the response over here and I will give the factor over here as marketing strategy and I will do the Kruskal Wallis test and I will click ok.

(Refer Slide Time: 15:07)

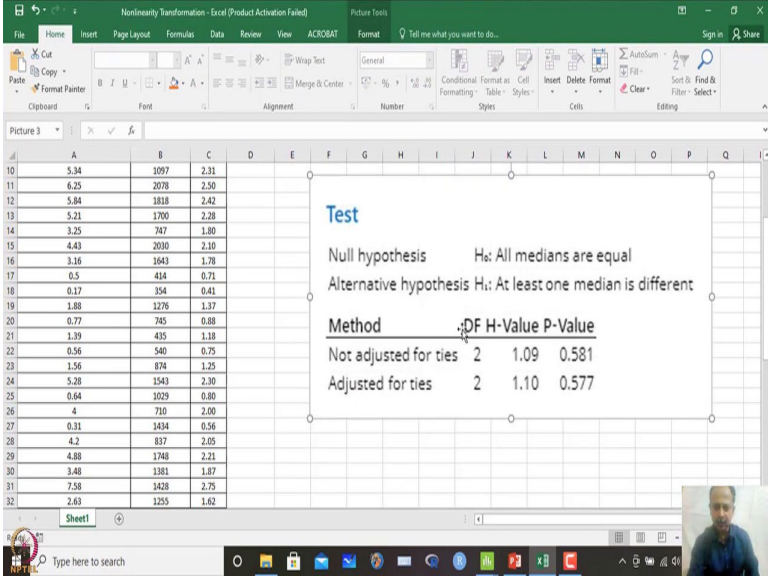


(Refer Slide Time: 15:09)



And then what I will get is that? This is one of table that we are concerned over here. So, I will copy this over here and I will just replace this and I will paste to over here.

(Refer Slide Time: 15:18)



The screenshot shows an Excel spreadsheet with columns A, B, and C. Column A contains values from 5.34 to 2.63, column B contains values from 1997 to 1255, and column C contains values from 2.31 to 1.62. A text box is overlaid on the spreadsheet, displaying the following information:

**Test**

Null hypothesis  $H_0$ : All medians are equal  
Alternative hypothesis  $H_a$ : At least one median is different

**Method**      **DF** **H-Value** **P-Value**

Not adjusted for ties	2	1.09	0.581
Adjusted for ties	2	1.10	0.577

So, when I do that what you see is that there is p values and it is using that median information not mean information, whenever it is nonparametric, they will use median information. Statistician suggests median information comparing and based on rank information that comparison test will be done and in this case whether the group median is same or medians are different at least one median is different.

So, in this case there will be two methods adjusted for ties and not adjusted for ties. We will go for not adjusted for ties, values of p values over here there is more conservative so we will go by that. So, 0.581 means that the medians are not different over here.

(Refer Slide Time: 16:18)

**ONE-WAY ANOVA AND ANCOVA.MXMX**

**Kruskal-Wallis Test: Car Sales (Error Non Normal)**

Design_Flexibility	Wavelength	Transmission_ANCOVA	Mark
Design Flexibility	12	6	16.0
Price	12	7	19.2
Quality	12	7	20.3
Overall	36		18.5

**Test**

Null hypothesis  $H_0$ : All medians are equal  
 Alternative hypothesis  $H_a$ : At least one median is different

**Method** **Df** **H-Value** **P-Value**  
 Not adjusted for ties 2 1.09 0.581  
 Adjusted for ties 2 1.10 0.577

**One-Way Analysis of Variance**

Response: Quality Design Flexibility Price

Options... Comparison... Graphs...

Results... Storage...

Background Spreadsheet Data:

C14-T	C15	C16	C17
INK	Wavelength	% Transmission_ANCOVA	Mark
1	A	698	47.16
2	A	705	52.83
3	A	712	58.63
4	B	695	38.47
5	B	702	45.36
6	R	708	52.56

(Refer Slide Time: 16:25)

The screenshot shows a Microsoft Excel spreadsheet with the following content:

**File -> Edit -> Data -> Calc -> Stat -> Graph -> View -> Help -> Assistant -> Additional Tools**

**ONE-WAY ANOVA: Quality...**

**One-way ANOVA: Quality, Design Flexibility, Price**

**Factor Levels Values**

Factor 3 Quality, Design Flexibility, Price

**Analysis of Variance**

Source	Df	Adj SS	Adj MS	F-Value	P-Value
Factor	2	18.06	9.028	0.68	0.51
Error	33	438.50	13.288		
Total	35	456.56			

**Model Summary**

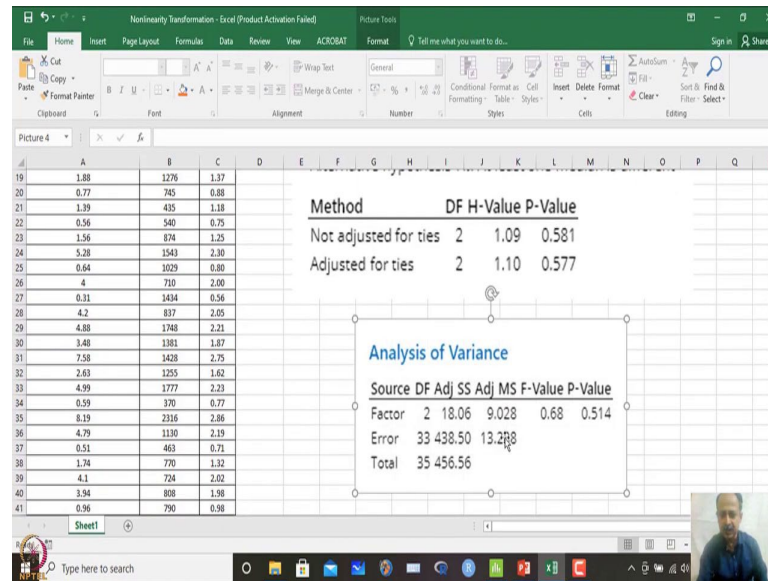
	S	R-sq	R-sq(Adj)	R-sq(Pred)
	3.64525	3.95%	0.00%	0.00%

**One-way ANOVA and ANCOVA:mxm**

	C14-T	C15	C16	C17	C18-T	C19	C20	C21	C22	C23	C24	C25	C26	C27	C28
	INK	Weightlight	% Transmission	ANCOVA	Marketing Strategy	Car Sales (Error Not Normal)		DC	DC+MK	MC		Quality	Design Flexibility	Price	
1	A	698		47.16	Quality		3	118.8	105.4	102.1		3	2	4	
2	A	705		52.83	Quality		6	122.6	101.1	105.8		5	5	2	
3	A	712		58.63	Quality		7	115.6	102.7	99.6		7	5	5	
4	B	695		38.47	Quality		4	113.6	97.1	102.7		4	3	6	
5	B	702		45.36	Quality		6	119.3	101.9	98.8		6	7		
6	B	708		52.56	Quality		7	115.9	98.9	100.0		7	8		

So, one way analysis of variance is also possible over here. So, what I will do is the data is in different column so, in this case quality design and flexibility. So, classical way if we have done this one also what we have seen is that classical way when we have done this one also, the p value what we are getting over here. If I copy this one classical approach for what we are assuming everything is going fine.

(Refer Slide Time: 16:40)



So, in this case what we are getting so, I can place this one and what we see is that p value is 0.514. Here with not adjusted for ties is 0.581. So both the analysis Kruskal Wallis at this one is giving me more or less same p value. Even if the distribution does not satisfy what we are saying is that analysis of variance is so, robust the conclusion made by non parametric is same as conclusion done by analysis of variance test that is ANOVA.

Even if the final condition or the final test had a model adequacy is not satisfactory then also I am seeing the interpretation comes out to be same ok. So, most of the time we can expect that it is so robust and it can be adopted like that. So, that is the way we should try to adopt these techniques like that and this is all I wanted to discuss about one way analysis of variance. So, we have discussed about all model adequacies and what scenarios we are adopting we are trying to screen the factors like that like from cause and effect diagram some of this factors like that.

And there is another way of screening the variables like this is although it cannot be guaranteed whether the factor influences CTQ or not. But preliminary some analysis can be done and which are the potential factors which I can isolate. Because you may not have done experimentation which is statistical experimentation, but you have some previous data historic data from where you try to interpret whether the factor influences the outcomes or not or CTQs like that. So, for that one of the important technique that we

will discussed briefly over here is known as regression. We will talk about only linear regression over here which is the primary aspect that is adopted in design of experiments and that is extensively used is design of experiments.

After even after doing experimentation we will use regression for developing the function between Y and X and from where we will go to the global optimal solution there from there we will go to the global optimal solution. But that is the primary idea that is required which I think is necessary over here to illustrate. So, what we are doing is that we are trying this is a conversion from control phase to improvement phase.

So, we are just in the border line over here and we are trying to see the potential factors and if I have screen the potential factors. And in that case I will do for full experimentation and from there I will identify which factor influences why and how much it influences and based on that we can optimize we can optimize the system or process like that ok.

(Refer Slide Time: 19:33)

Quality Control and Improvement using MINITAB

**Linear Process Modelling**

**Typical Questions:**

Ticket sales in a football match?

Change in height or weight per unit time period for a baby?

Product Demand next week?

Daily Room booking in a hotel ?

Output of a Process for given inputs and process control variable settings?

$Y = f(X)$  — OPTIMIZE

-> **Regression Analysis**: a statistical technique for modeling the relationship between variables.

**Linear regression** is a general approach for estimating/describing association between a **continuous** outcome/response variable (dependent) and one or multiple predictors.

- **One predictor**: Simple linear regression
- **Multiple predictors**: Multiple linear regression

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

So, next important topics what we want to discuss is regression important topics that we want to discuss. So, our overall objective is to develop this mathematical function over here because if I can develop the function, I can optimize this function over here.

So, this mathematical function and there is this regressions technique is used whenever mathematical model does not exist, that means, physical model does not exist or

mechanistic model does not exist then only we go for it. And this is true because machines are working for many years and we will not find that the previous when it was installed the scenario is like that. Due to wear and tear what we do is that we tried to develop new models like that.

So, it is stochastic that we want to develop at a given time point what is the scenario and what is the mathematical relationship and based on that we will adopt optimization and try to optimize that one. So, in this case this is empirical modelling what we call regression analysis this modelling, but this is empirical relationship that we want to establish between Y and X over here ok.

So, this types of models are extensive not only in design of experiments or in processes you can find out the applications of these in football matches that how much ticket we will we can sell, based on the different conditions of the match like that who is playing and weather conditions and all the scenario where it is being held and all these things.

We will dictate how much ticket will be sold like that. So, there will be some predictors and there will be some outcomes or CTQs that we want to predict over here. So, this is a prediction model basically what we are adopting over here and in and this is also used in design of experiment.

So, then also we can see what should be the change in heights or weights. So, change in heights or weights per unit time like that. So, these things can be we can predict, demand like in operations what we do is that we want to understand demand for car what will be the demand like that for next period like that and for that what we do is that we adopt this regression techniques also to extrapolate and try to see that what will happen the at  $t + 1$  condition like that.

So, I have information up to  $t$  what will happen  $t + 1$  like that ok. So, room booking in service industry also we can think of that when we are when we are trying to predict something so regression equation can be used. So, similarly CTQs of a process can also be we want to see and for that this is the Y and we want to see how it is influenced by different factors, which is X over here.

So, X can be  $X_p$  variable that is here or  $X_z$  variable over here, that we have talked about and over here, and it can also be covariates that can influence by process like that



what we are discussing just now what we have discussed. So, there can be variables which I can control there can be variables which I cannot control there can be covariates or input conditions which also keeps on changing and that can influence the CTQs.

So, regression can also be used to understand the relationship between this. Why I want this relationship because I want to optimize that also I want to optimize the total process for that I will use some optimization techniques. So, over here to reach to the optimal scenarios where or what should the X condition that will basically optimize by Y ok.

So, this is one of the easiest techniques that we will learn over here is linear regression over here. So, and also the simplest one is I have one predictor over here that is one X variables over here and one predictor. That means, one Y and one X scenario that is known as simple linear regression that we are trying to understand over here.

I will not go into complexities of many other complexities of regression, but I will say what are the scenario and how we can apply that in MINITAB interface

(Refer Slide Time: 23:09)

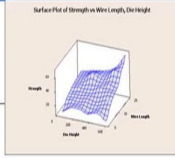
**Quality Control and Improvement using MINITAB**


(Y)

- Dependent variable (s)
- Output (s)
- Effect (s)
- Symptom
- Monitor
- Response(s)
- Predicted


✓  
 $X = X_1 \dots X_N$

- Independent variables
- Inputs, and In-Process Variables
- Causes
- Problem
- Control
- Inputs and Process Setting Conditions (X)
- Predictors





**Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay**



So, some theories behind this which we have we will try to understand and then we will adopt that one and apply that one in MINITAB when we will try to interpret the results like that ok. So, this is Y variable and these are the X variables X can be N number of variables over here.

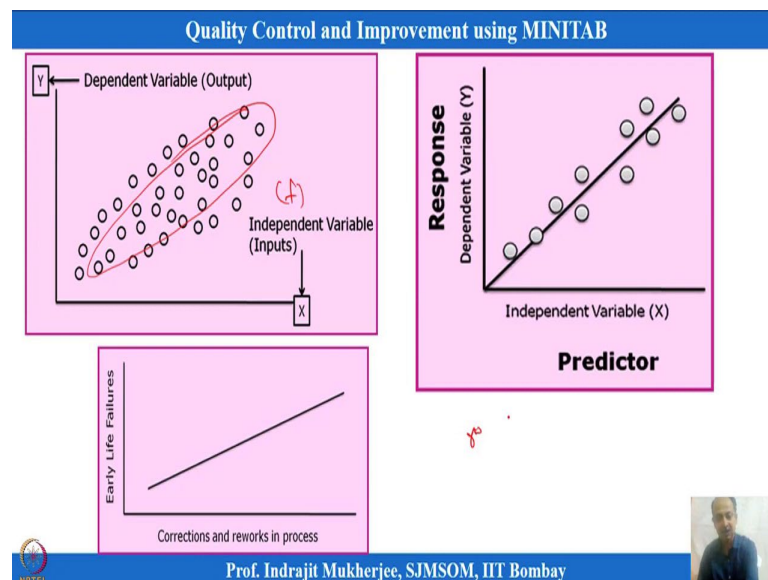


So, this is known as independent variable on this side X is known as independent variable and Y is known as dependent variable. This can be known as output this can be known as input over here, this can be a cause an effect this can be effects this is known as cause over here.

This is the symptom and this is the problems, this is what we are monitoring over here and what we control over here that is X variables that we control over here. This is also known as response. MINITAB understands response over here and there can be inputs conditions like that process setting condition that are X variables that we are telling. It can be predicted this is known as predicted and this is predictor basically predictors we can think off ok.

There are different names for Y different names for X over here. So, you can think about cause and effect. So, one is cause one is effect basically and so, you have to; you have to selective which I can control is X which I cannot control is basically Y. So, that is the interpretation we can make out of this ok.

(Refer Slide Time: 24:18)

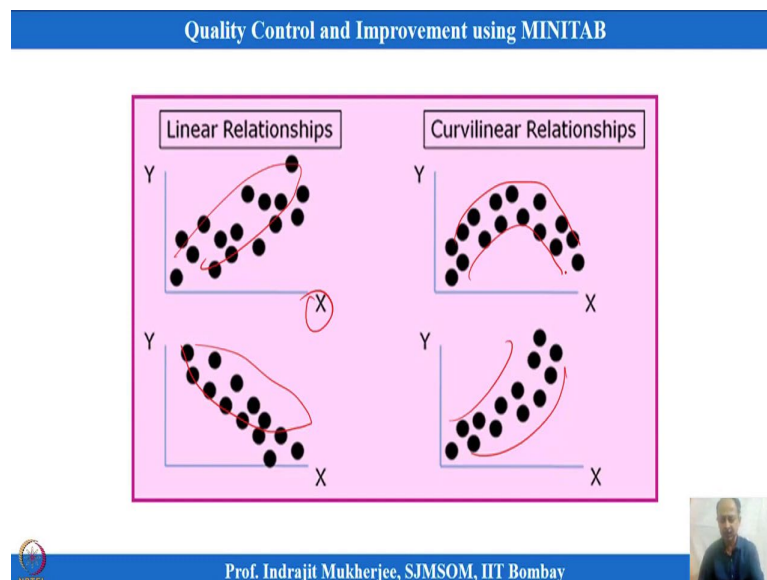


So, what we will do is that this is Y and X like that and for that preliminary what we have discussed is the correlation coefficient and scatter diagram that we have seen in visualization of the data like that. What we have used is that whether the relationship is linear or not so, this kind of relationship whether it is positive or not.

So, for that what we used is correlation coefficient that is important correlation coefficient. So, to develop regression equation we have to first see the scatter plot and try to figure out that whether the relationship can be linear and then adopt the regression model.

And how this model is developed that is important some theoretical concepts over there how this models are developed. And based on that we can go ahead and a many of the books and a videos will explain all this more theories about how these models are developed basically.

(Refer Slide Time: 25:06)



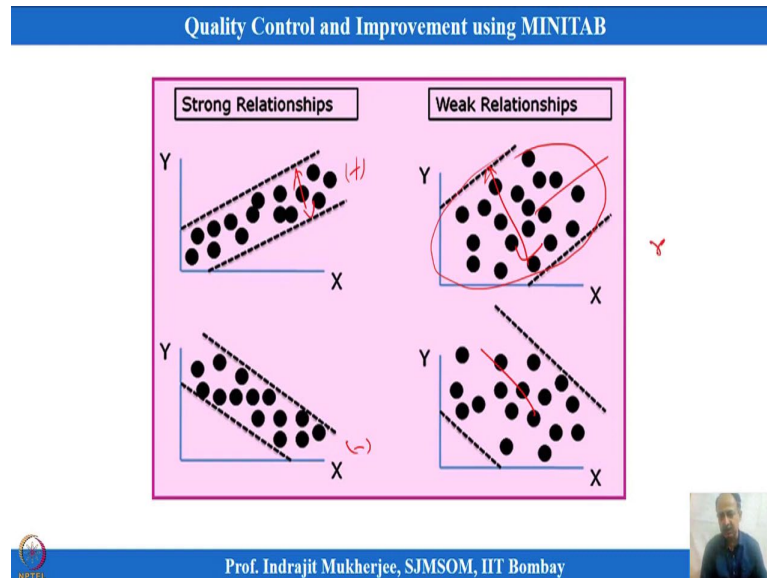
We will take the simplest model over here, one single  $X$  and one single  $Y$  over here. So, it can have positive relationship what we can see this can have an negative relationship what we can see ok. There can be non linear relationship that you curvature that you are seeing over here this is the curvature that you are seeing over here.

So, this we can think of polynomial equations over here or non linear relationship that exist between  $X$  and  $Y$ . There can be different types of relationship that we want to understand so, but scatter plot is important over here.

So, if you have single  $X$  and single  $Y$ , I can plot that in scatter plot and see the relationship and based on that I can understand that whether a linear model or a non-

linear or a polynomial equation will work over here, and based on that we will adopt that type specific models like that ok.

(Refer Slide Time: 25:48)



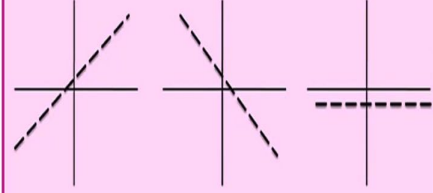
And if the X and Y are these are the shapes over here this we can say that there is a positive relationship that exist. If this is the scenario this is negative relationship this is not so strong relationship because the variability because you see the width of this is much less as compare to the width of this over here.

So, width is more means relationship is weak and width is very small and all the data's are confined into a small tubes like that if you can think of. So, in this case what happens is that relationship is more strong and this is a weak negative relationship, this is weak positive relationship or we can see and the measures that we will use this correlation coefficient to understand this one.

(Refer Slide Time: 26:24)

**Quality Control and Improvement using MINITAB**


**Types of Correlation**




Positive Correlation
Negative Correlation
No Correlation

Sample covariance =  $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Sample coefficient of correlation:  $r = \frac{s_{xy}}{s_x s_y}$  cov = r



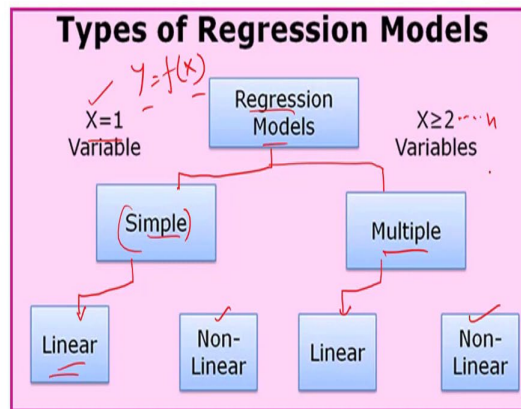
Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay



And correlation is nothing, but the covariance between x and y that we can calculate and all the software's will give you the covariance information and because covariance is not bounded over here, what we do is that we take correlation coefficient where we divide it by standard deviation sample standard deviation.

So, covariance by sample standard deviation, so standard deviation of x standard deviation of y that will give me the r sample correlation. And this can we need have also checks this one by p interpretation like that and t test is use for that and it will give you the p interpretation also. So, MINITAB gives you whether the correlation is significant or not significant like that using hypothesis concept like that ok.

(Refer Slide Time: 27:07)



So, there are different types of regression models over here. So, I am trying to understand first simple regression simple linear regression over here, and the this is the one that we wanted to stand. So, this I go from this regression model I go the simple model and this is the linear model over here.

So, a similarly multiple regression we can see that linear models how it is develop like that. We will skip that non-linear relationship over here. So, whenever I have a X which is the predictor, I have a single Y and function of X we want to develop. If we have multiple X, in that case what will happen is that this is up to n number of variables let us say to generalize this one and this known as multiple regression over here.

(Refer Slide Time: 28:04)

**Quality Control and Improvement using MINITAB**

**Simple Regression Model**

$y = f(x) = \beta_0 + \beta_1 x$   $y = mx + c$   $f(x) = f(x)$

$X = \text{Continuous}$   $Y = \text{Expected Response}$  Slope Coefficient Independent Variable Error in model estimation

$$E(y|x) = \underbrace{(\hat{\beta}_0 + \hat{\beta}_1(x))}_{\text{Linear component}} + \underbrace{(\epsilon)}_{\text{Random Error}}$$

$y$   $x$

Simple regression considers a single **regressor or a predictor** (x) and a **predicted or response variable** (y)

Prof. Indrajit Mukherjee, SJMSOM, IIT Bombay

So, how regression works? How to estimate the coefficients over here, this was given by Gauss. This was developed very long back like that about 1900 approximately at the time point maybe also we can check that one. So, any how this is highly useful techniques and what it tries to say is that I can develop a mathematical function which is between expected response over here. And I want to develop the mathematical model.

For given condition of X and Y, I have different observations over here ok. And based on the many observations over here, how can I draw a functional relationship or functional linear relationship over here that is important over here. So, there are many points and I know to develop a line equation only two point is sufficient.

So, if you have two points I can develop what is the line equation over here and in that case what is the slope over here and what is the intercept that we can calculate over here. So,  $mx + c$  that can be calculated based on equation over here. Similar approach is taken over here also, but the only thing is that there is no two points. Here are multiple points like that and I can have multiple lines over here. n number of points we have and any two points I can take.

So, " $C_2$ " is a combination that I can think of, but one of the best lines I have to adopt out of this. So, which is the best line I should adopt which will explain the relationship between Y and X. So, I want to develop a line equation which is best fit and for that some theoretical aspect is taken over here which is which Gauss has suggested and he

says that expected value this so, whenever I have a regression of X and Y. So, each of these value at a given point of X, what will happen is that I can have multiple over here, like ANOVA analysis what we have seen is that for a given scenario. If I change the condition over here and reset that one what will happen is that I can get multiple values of Y over here.

So, there will be some mean values and there will be some variations over here. Similarly, at a given different points there will be variations like that. So, idea of analysis of variance can also be extended over here in regression, only thing is that in regression what we are considering at this stage is X is continuous also and Y is also continuous Y is also continuous over here so ok. So, X is continuous Y is continuous.

Earlier X was discrete and Y was continuous over here, what we are considering is X is continuous and Y is continuous over here. So, how this line equation is developed? That is of importance to us. So, over here what it says is that it can be expressed as like intercept what we have seen. So, this we can think of as  $c$  intercept over here, this we can think of a slope like that. So, interpretation remains like line equation like that. And the one is known as  $\beta_0$  and one is known as  $\beta_1$  over here. These are the two important parameters we want to estimate from all the points that we are getting X and Y over here and that is the function that we want to develop. So,  $f(x) = \beta_0 + \beta_1 x$  and that is the approximation and this is the function of X that we want to generate over here from the given dataset which can be n number of observations like that.

So, this we will continue discussion on this and we will stop over here we will start from here. So, this is the basic idea. So, what we are talking about is that we are trying to develop a linear models and simple linear regression one simple Y CTQ and one X over here and we have a historic data point no experimentation over here.

I want to identify whether the X influences Y over here, although causal relationship cannot be established by regression, but at least some hints of potential whether I can consider for further experimentation or not that some hints can be we can get out of this ok for that we use a regressions if I have some previous data like that.

And here we are considering X is continuous and Y is continuous just extension of ANOVA analysis it is more generalized you can think off. So, ANOVA is at the discrete

X variable X points like that X levels over here X can have any values like that it can be continuous.

So, this condition, so expected value of Y for a given X is equals to  $\beta_0 + \beta_1 x + \epsilon$  with some error ( $\epsilon$ ) over here, every model we will have some error I cannot exactly model expectation of Y over here for given X. I cannot have a perfect function like that every function will have some error; that means, I will go wrong, but I want to reduce that minimize that error like that. So, I want a function which is very close to the reality, so that I commit minimum error.

But there will be some error we cannot avoid that one because this is the empirical relationship and this is and the it depends on the scenario of the machine or scenario of the process like that. So, it cannot be exactly model because I will miss out some factor.

So, I cannot be exactly close like that, but there will be some error always there will be some error when I am developing this mathematical function using regression like that ok. So, this we will continue from here ok in our next session we will continue from here in our next session.

Thank you for listening.