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

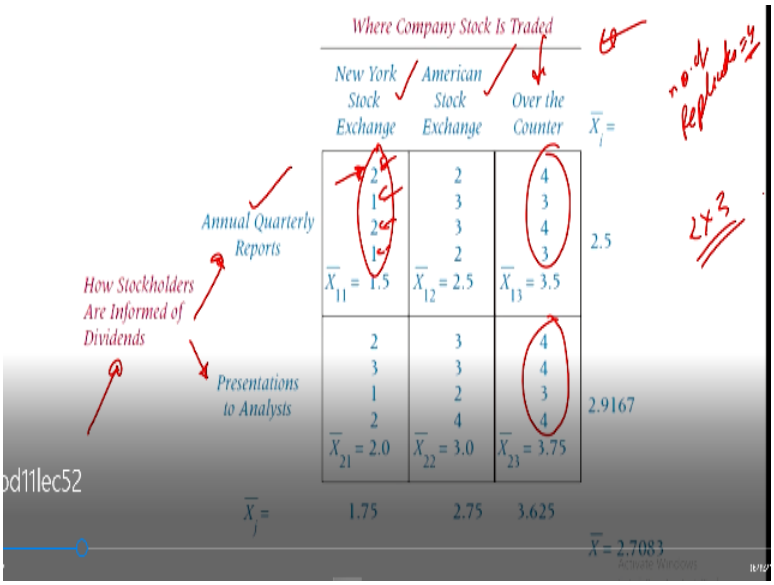
Lecture – 52
A Factorial Design - II

Hello friends, I welcome you all in this session as you are aware in previous session we were discussing about factorial design. Factorial design is a kind of experimentation technique wherein we analyse the effect of 2 or more than 2 independent variables on dependent variables. So how factorial design is different from other 2 designs which we have seen the first experimental design was completely randomized design.

Wherein we have seen the effect of 1 independent variable on a dependent variable the second 1 was randomized block design where we have seen the effect of again 1 independent variable on the dependent variable. Even that there was 1 blocking variable so let us take an example of 2 independent variable and a dependent variable. So if we take 2 independent variables we can solve this example by completely randomized design experiment as we will.

But the point is we will have to perform completely randomized design experiment 2 times because you will have to take 2 independent variables separately. So this is one way the second way could be a you take 1st independent variable and the 2nd independent variable can be blocking variable or the 3rd one is factorial design you just take both the independent variables simultaneously and see how they are affecting dependent variable. So you would be aware of the previous session we were discussing about this question.

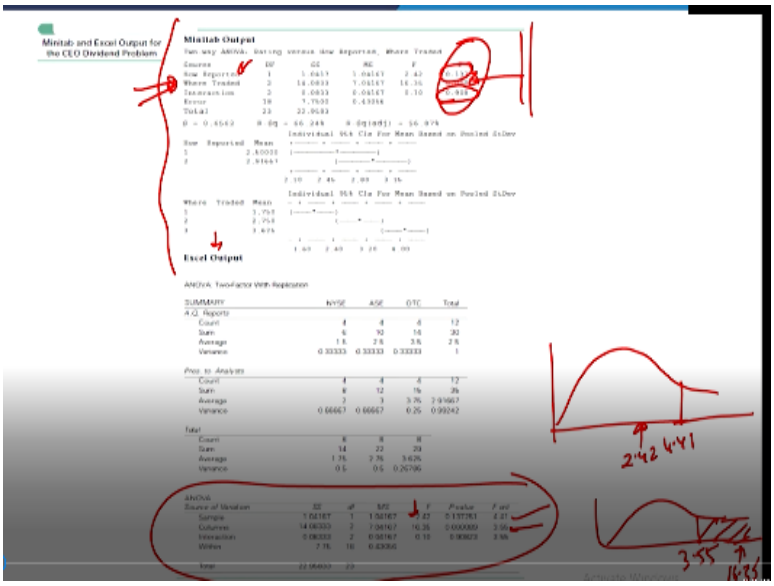
(Refer Slide Time: 02:27)



So we want to find out the effect on the stock price on when there are 2 independent variables the first is where the stock is treated, traded and how stockholder are stakeholders are informed about dividends. So company can trade stocks in these 3 you know stock markets let us say New York stock exchange American or over the counter this is the 3rd one. So here how you know informing stockholders either through annual or quarterly report or through presentations right.

So we have taken data 4 times of 1 particular stock so there are we can say that there are 4 replicates right so the number of replicates are number of replicates are 4 right. So the value of stock in New York stock exchange when the reporting system was this annual and this was the value of the stock then second trading 3rd and 4th. Similarly, these 4 and so on right so this is the case of 2*3 design right its a you got let us say 2 levels of this independent variable and 3 levels of second independent variable.

(Refer Slide Time: 04:02)



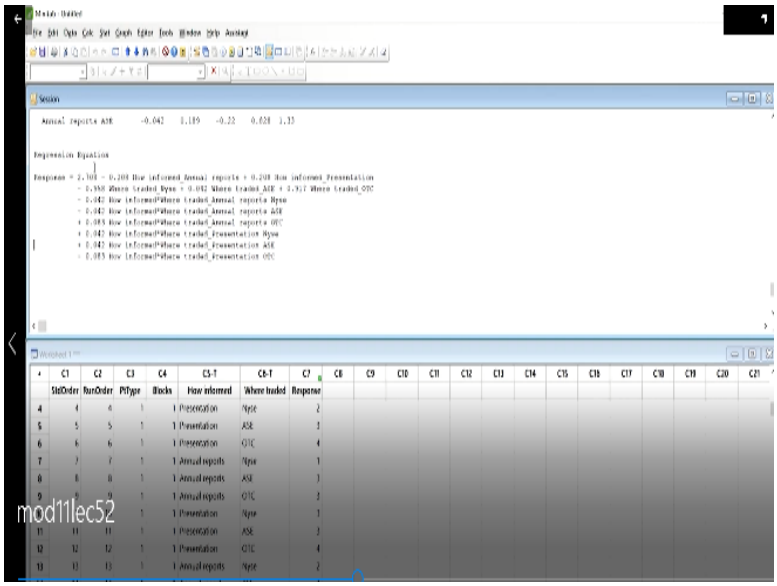
So we have seen these output as we will in previous session so after calculating all unnecessary statistics let us look at Minitab output so this is P value right so this values as < 0.05 so we will say that there is a significant difference in stocks as far as where the stocks are traded right but if you look at p value for this how the reporting system was. So there is no significant difference so whether the stakeholders are informed through presentation or through annual reports.

There was no significant difference between these two. And of course interaction is again its not there is no interaction effect so there is only one conclusion that there is a significant difference are the independent variable this where the where companies stock is traded is affecting dependent variable right. So this is again output from excel this excel output so you can see here F critical is this and F observed is this right.

So it is how you would be drawing your distribution so F critical is 4.41 and your f observed is this let us say 2.42. So you are not rejecting an hypothesis and you are saying that the reporting system is same in both the methods. But if you look at the second one this one right whereFf critical is where F critical is this is 3.55 and calculated value is 16.35 right. So this is a rejection region so you are rejecting the line this state.

So we will work out this example using Minitab and you need to enter data very carefully and then we will see we will analyse output right.

(Refer Slide Time: 06:21)



So let us look at how to enter data so go to stat we will go to design of experiment and then factorial design create factorial design right now this is general full factorial design. So we know that the number of factors are 2 because the first is how stocks are traded and how the stakeholders or stockholders are informed right. So number of factors 2, go to design so you will get this design right.

So A and B so each has got 2 levels right so the first has got 2 but the second has got 3 levels right and when we say B, B means how you are trading stocks in different stock exchanges right so this is your number of levels for A and B okay so this is we have got 4 replicates right as I said we are taking observation for 4 continuous days right so number of replicates 4 click okay and go to factors.

So you can name them let us say how stocks or how stakeholders are informed. So it is of course a non-numerical one there it is annual reporting are through presentation okay then for B where it is traded where stocks are traded all right. So this again text I will be willing to click on this arrow as text so are 3 options now so it is; let us say new NYS New York NYSE New York and second is write the other one and the 3rd one was OTC over the counter OTC right.

Click OK go to options we will have it is up to you if you wish you can have some data store somewhere then OK. Of course how you want to have printout whether all the graphs in just one page or separate page for each graph you click OK. So this is how you should enter data right so there are total 24 data points which you need to enter right so this is response variable now you have to be very careful right.

So annual New York Stock Exchange response 2 annual the other stock exchange is 2 over the counter 4 then presentation New York Stock Exchange presentation the American Stock Exchange and presentation 4 OTC. So this how you are supposed to do data entry in Minitab. So this is now annual report New York Stock Exchange annual ASE annual OTC then presentation ASE OTC annual report New York Stock Exchange, American stock exchange and then OTC.

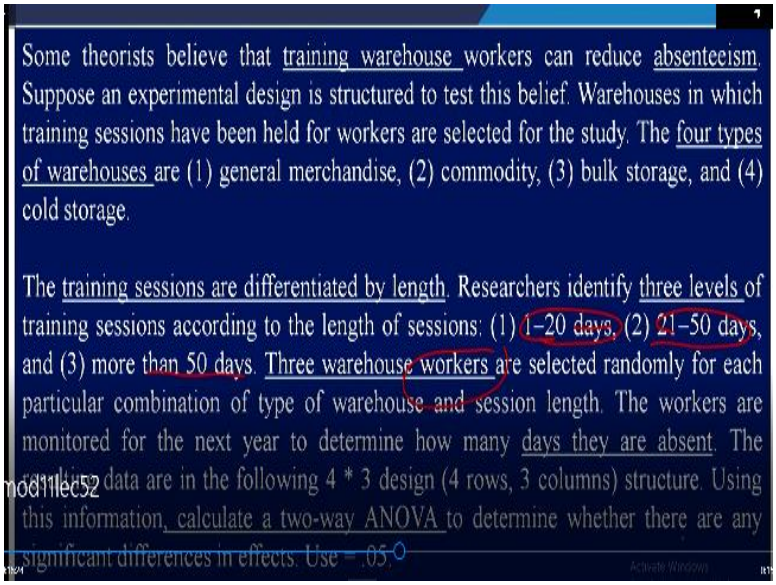
Again presentation and this final data point which you are supposed to enter right so this is your response variable when you go to click design of experiment, factorial design, analyse factorial design right. So response variable is to be selected so select response variable you can enter the significance level which is 95 let it be like this let all these be there. So here you will get lots of output we have to just look at p value right yeah.

Just see this so p value is 0.137 how they are informed so there is no significant difference between whether stock holders are being informed through presentation are through annular quarterly report. But there is a significant difference and just see this so far as where stocks are being traded whether it is New York or American Stock Exchange or over the counter right. So this is how in fact interaction if you look at it there is no interaction.

Because this value is more than 0.05 so there is no interaction effect only there is a significant difference amongst those 3 columns right. So this is what you can get as an output you can also get F values and of course there are other set of information you have got R square and adjusted R square and so on all these things are not needed at this point in time in fact you are getting a regression equation as well.

So this is your constant term 2.70 so how you know stockholders are informed whether it is through annual report or through a presentation and where it is traded. So this is a complete multiple regression equation so we will not see this at this point in time so this is the way you should solve equation 2 way ANOVA. Right there are 2 independent variables and we are seeing the effect of both of them simultaneously on dependent variable. So let us move on to next question which is also on two way ANOVA.

(Refer Slide Time: 14:37)

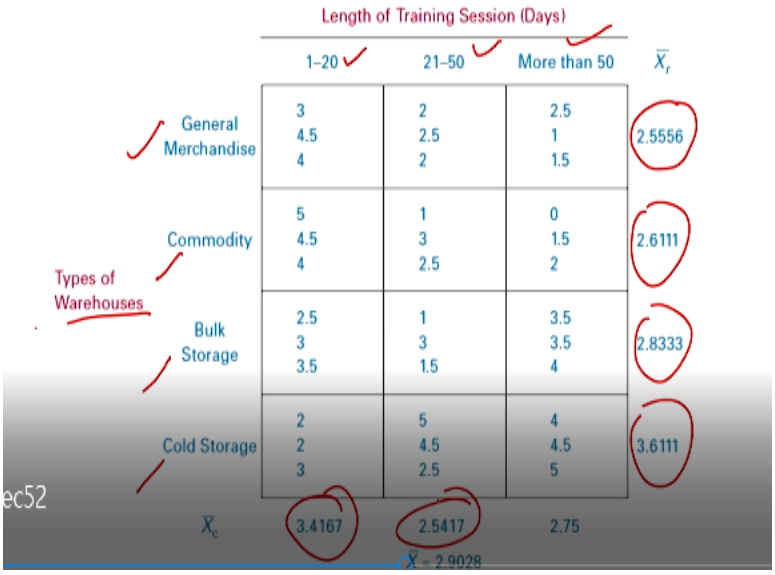


So let us say some theories are let us say managers believe that the training to the employees of the warehouse will reduce absenteeism right and they thought of doing some experiment. So they selected a workers from 4 different ware houses. So let us say warehouse number 1 general merchandise warehouse second is commodity warehouse bulk storage warehouse and 4th is cold storage warehouse correct and there were different training sessions by length.

So there were 3 different sessions so you have got 1st session of let us say up to 0 let us say 1 to 20 days 21 to 50 days and more than 30 days right sorry 50 days then after selecting 4 warehouses and 3 different training sessions 3 workers were selected from each of these warehouses so they were put on training the workers of these 4 stockholders they were put on trainings on these 3 different training sessions.

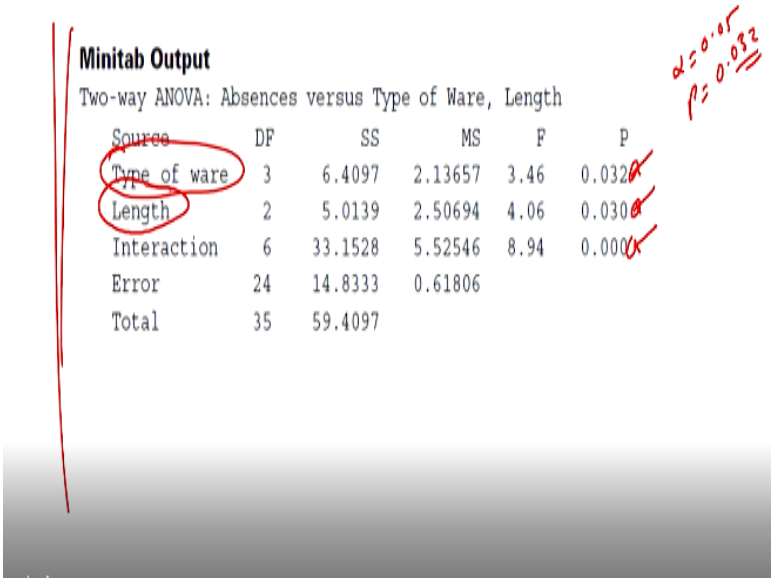
And their absenteeism was noted right so we so we wanted to know whether there is a significant difference in absenteeism amongst types of warehouses workers of types of ware houses and going through different training sessions right. So we have to calculate two way ANOVA to determine whether there are any significant differences in effects. So is absenteeism depends on these types of warehouses and length of training sessions right.

(Refer Slide Time: 16:44)



So this is the data which was collected again here there are different readings. So length of training sessions 1st session 2nd session 3rd session and the types of warehouses general merchandise commodity, bulk storage, cold storage. So you have got this averages over here right of all these 4 rows all these 4 columns right and this is the overall average right.

(Refer Slide Time: 17:24)



Minitab Output

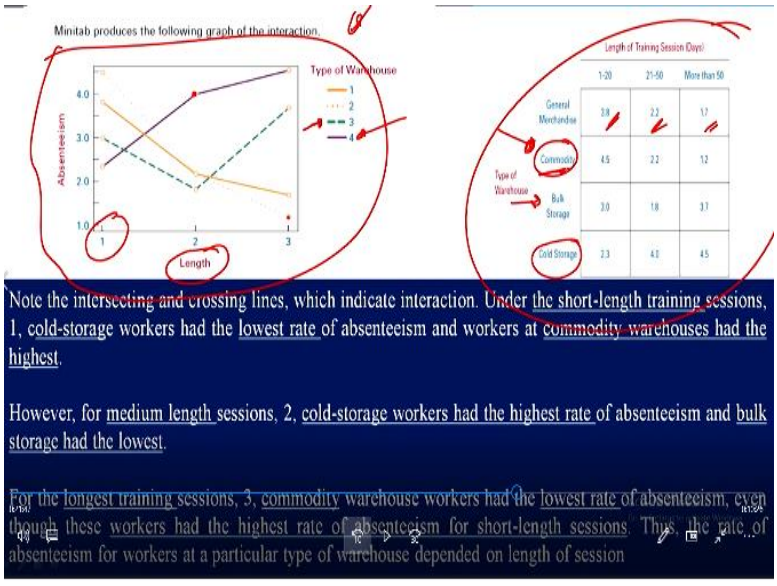
Two-way ANOVA: Absences versus Type of Ware, Length

Source	DF	SS	MS	F	P
Type of ware	3	6.4097	2.13657	3.46	0.032
Length	2	5.0139	2.50694	4.06	0.030
Interaction	6	33.1528	5.52546	8.94	0.000
Error	24	14.8333	0.61806		
Total	35	59.4097			

So let us look at the solution to this question so this is the Minitab output which you would be getting it right so if you look at the significance level which I think is given in the question it is 0.05 so p value so alpha is p value here 0.032 p value is < alpha so we will reject null hypothesis again we will reject null hypothesis and for interaction effect again and we will reject null hypothesis it means that the absenteeism is different in different types of warehouses and it is again it depends on different sessions and there is an interaction effect right.

So let us solve this question and before we solve this question.

(Refer Slide Time: 18:23)



Let us look at the interaction result or the output from this particular software. So this is this basically this graph this plot is interaction effect. So we know that this is the mean of all these 3 values which we have seen earlier as well. So mean of these 3 is 3.8 2.2 1.7 right similarly for all other sales right now I feel if you look at this plot carefully so when the training session is short length then absenteeism is least for 4th type of warehouses.

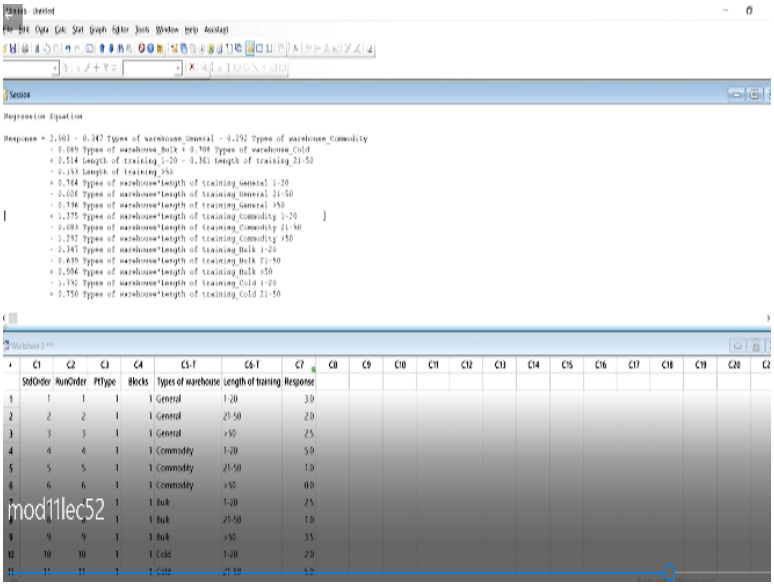
The workers of 4th type of warehouse 4th is what cold storage so we will say that when the training session is of short duration then the absenteeism is lowest for the workers who are working in cold storage warehouse and highest absenteeism is where this series this dotted point right. So this is warehouse number 2 it means commodity warehouse. So at a short length training session the highest absenteeism is for commodity stock warehouse workers and least is for this.

But when the training session is off medium duration then the lowest absenteeism is for this 3rd type of warehouse right and this is bulk storage. So for medium length sessions which is session number 2 cold storage workers had the highest rate of absenteeism just see highest rate is here right it means 4th one right. So highest number of absenteeism is for the is amongst the workers of commodities warehouse.

But for medium damn session the highest absenteeism is for cold storage workers and what about this when the training session is the longest one the absenteeism is lowest for commodity warehouse right. Now if you look at this the absenteeism is lowest here for commodity warehouse workers but highest when training session was of lower duration. So we will say that the absenteeism depends on the workers of different storage warehouses.

And it also depends on length of training session and there is interaction amongst these 3 these 2 right.

(Refer Slide Time: 22:22)



So let us look at how to solve this question using Minitab so when you can delete this output first because it would be easier for us to see the output for second question. So just control and delete right now we will have data entry for our second question so stat DOE factorial design create factorial design first so general full factorial design. So a number of factors are 2 right then you have designs.

So here you can write number of levels first so there are 4 types of warehouses right. So this can be types of warehouse and the other one is training session length of training session length of training will also suffice so it has got 3 levels and the first 1 has got 4 levels right number of replicates 3 right so we will go to OK then factor right text so here you can write down write down the names of warehouse general, commodity, bulk and cold storage right.

So here you can have different training sessions let us say 1 to 21 21 to 50 and more than 50 right click OK yeah this is how you have designed your experiment and now the point is you have to write your response variable right. So this is your response or absenteeism right so first warehouse general length of training session the shortest duration then medium duration longest duration similarly for commodity type of warehouse.

For bulk then for 4th one which is cold storage right then general this is how you are supposed to go for data entry as I said if you have got large number of factors and levels try to reduce number

of replicates right because you would be otherwise consuming lots of time and efforts. So this how you are supposed to enter data right now you can analyse this question DOE factorial analyse factorial design right.

So response variable is C7 options so you can use confidence level is 95 click okay. Let us look at p values for this question. So type of warehouse p value is 0.032 same as I will show you this 0.032 type of warehouse let us say 0.032 length of training session 0.030 and for interaction in this 0 right. So we will say that there is a significant difference as far as absenteeism is concerned of workers of different types of warehouses and the length of training sessions which they undergo and as well as there is an interaction effect as well right.

So this is how you should analyse output and again you will be having the multiple regression equation the R square adjusted R square and so on right.

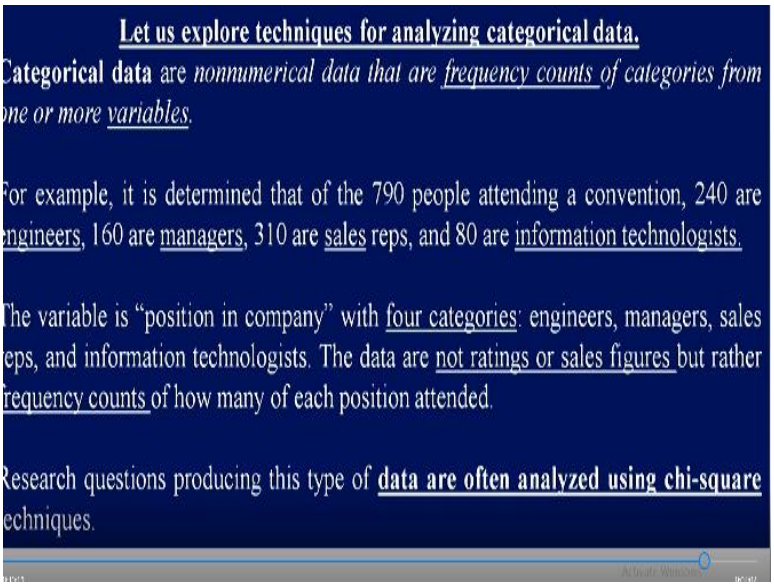
(Refer Slide Time: 28:04)

So with this let us move on to the next topic which is on Chi square test in fact we have seen several analyse analytical tools we have seen regression we have seen a coefficient of correlation we did not see regression so far we have seen correlation we have seen a coefficient of co variants we all seen ANOVA we have seen and in ANOVA we have 1 way ANOVA, 2 way ANOVA and we have seen factorial design.

But what when we were solving questions on hypothesis testing of means of 2 samples we have analysed not only hypothesis testing of a means of 2 samples but we have analysed variances as well we have analysed proportions as we will now sometimes in fact many times you may face a situation where you will have to an analyse proportions of more than 2 samples. Now in situation like that you cannot have just as simple as that tester to t test.

Because you have got more than 2 proportions of course you can always have 2 T test you just select first and second and proportions compare them then 2nd and 3rd and 1st and 3rd. In fact there would be 3 separate 2 test you will to perform but there is an efficient method available which can be used for analysing proportion of more than 2 samples. So this is a tool called Chi square test.

(Refer Slide Time: 30:11)



This tool is applicable for analysing get their medical data or non-numerical data so we collect data not in terms of let us say on ratio scale on rank order scale but we collect data on frequency counts right. So let us take an example let us say there are 790 people attending a convention 240 are engineers, 160 managers, 310 sales representatives ,80 are information technologists and so on.

So here what we have said and there is one variable which is a categorical variable and non-numeric variable and it has got 4 different categories right. So those categories are let us say

engineers, managers, sales representatives and information technologists right. So when we have got research questions I am producing such type of data they are analysed using a technique called Chi square test.

You will see in detail about Chi square tests for the time being let me stop here. We will work out couple of examples using Chi square test in next session. Thank you.