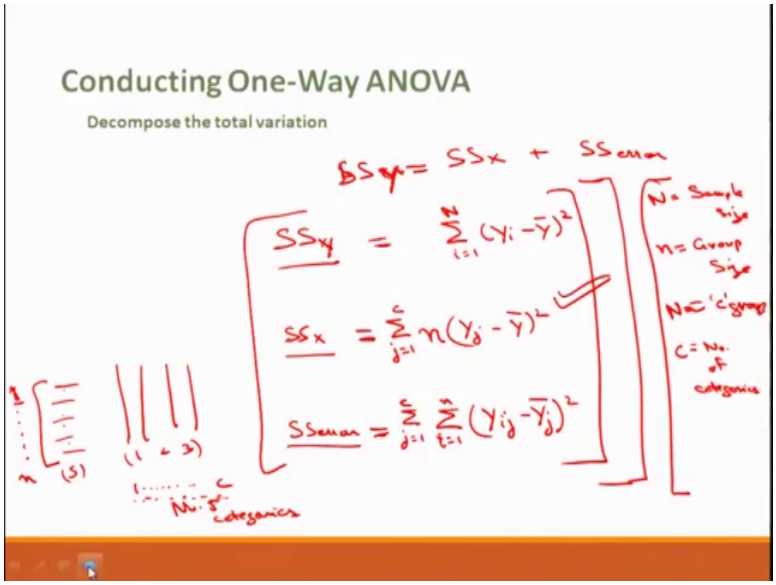
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 24
Analysis of Variance (ANOVA) – (Part-2)

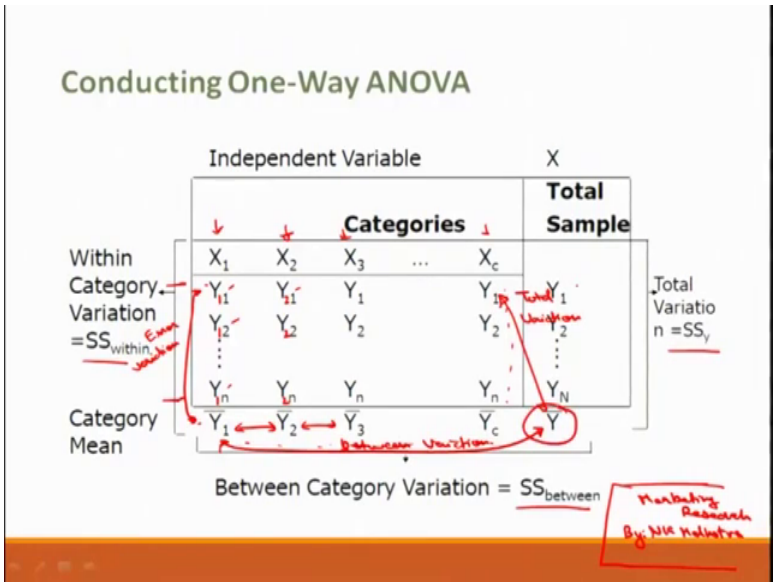
So, welcome to the part 2 of analysis of variance session.

(Refer Slide Time: 00:21)



So, where were we? We were discussing these relationships for total variation, between variation and error variation.

(Refer Slide Time: 00:33)



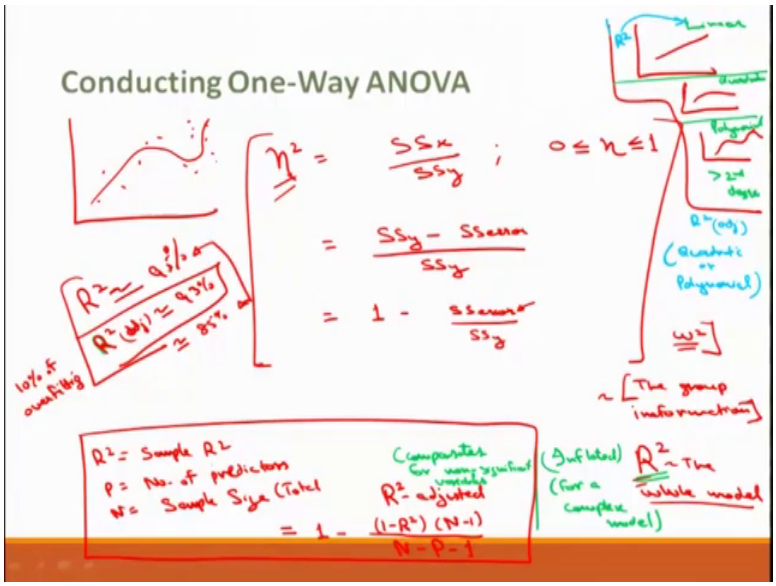
So, let me put it in a tabular form here. So, we have this table. So, this data is taken from the book by NK Malhotra. This illustration is taken from here it is marketing research the example which I will discuss here is also taken from the same book. So, you can read further if you like to have more information. So, what we have? We have these categories. So, mind it we had capital N here; capital N here is sample size not the population size.

So, small n is my group size; where capital N is divided into divided into C groups and therefore, C here is the number of categories. So, we have c categories 1 to C categories here 1 to n is my group size we can see the all group sizes are same here and we will see that the l variation here is between my element Y and Y 1 bar you can see this here the error Y ij minus Y j bar Yj bar is Y 1 bar this minus this is error variation ok.

So, between these between these implies between Y 1 bar and Y bar this is my between variation and for each element for Y 1 Y 2 shown up to y n and. So, on for category 2 Y 1 why I can even put this is x Y 1 1, Y 1 2, Y 1 n this can be Y 2 1 222 n for each element here for each Y i i varies from 1 to n here, we have the total variation I can put it here this is total variation. So, total variation is SS y between category variation is SS between and within category variation is SS within. So, we have this category means here and this is the total sample mean Y bar.

So, let us see how to conduct this ANOVA analysis using this table.

(Refer Slide Time: 04:16).



So, to conduct ANOVA we need to see the eta square. Eta square is SS x by SS y that is between variation over total variation, which varies from 0 to 1 this can also be put as SS x can be put as SS y minus SS error over SS y, we can say this 1 minus SS error by SS y. So, if you see this we have come SS error here. So, eta square and 1 more statistic known as omega square are more focused on the group information, it tells us about the within group variation and there is statistic known as R square which is nothing, but square of the correlation coefficient that tells us about the whole model.

So, R square and eta square are varying two different things, R square is the contribution of the entire model the whole model in explaining this study variation, eta square and omega square measure the contribution of individual model term. So, they are different eta square has a positive bias; So, it over estimates the individual contributions. So, from a practical stand point the bias is small. So, it depends upon your situation the data behavior how we have obtained that.

So, R square can be significantly inflated by adding non significant terms in the model. So, there is a term known as R square adjusted; R square adjusted is for compensating the non significant terms, which have inflated. So, we have inflation here in R square it is inflated for a complex model and R square adjusted compensates with the compensate note with compensate for, non significant terms.

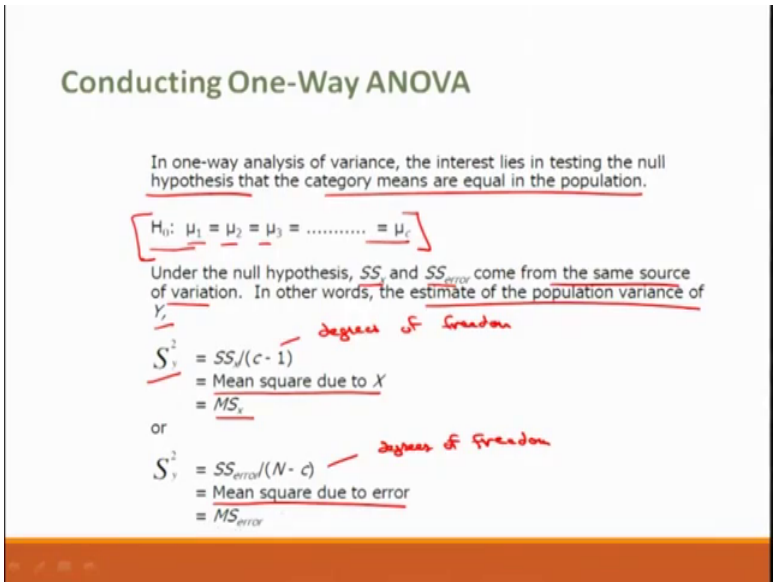
So, the R square value has a bias it over fits the model, there are chance correlations here that trends in (Refer Time: 07:49) and certain variable forms can also influence this one. So, R square adjusted is calculated as I will put it here, R square adjusted is 1 minus this ratio 1 minus R square into n minus 1, n minus p minus 1 where R square is the sample R square, p is the number of predictors and n is the sample size a sample size I mean that total sample size.

So, this is R square adjusted that gives the more close or more accurate information about the fit of the model. So, by fit of the model I mean if R square value is let me say if this is 95 percent, but it is over fitting memory it is its telling that the model is 95 percent close to the original data means for instance I have the data points here, and this is the model here. This is the kind of a cubical model the way I have drawn it.

So, we have various kinds of models, this is given linear model, this would be a quadratic model and this would be a cubical or some polynomial model linear quadratic and polynomial that is more than second degree more than second degree polymer. So, in this case if R square value is telling 95 percent value, R square adjusted value is telling R square adjusted value is somewhat 93 percent than this is more close or sometimes if R square adjusted value is may be 85 percent and this difference this 10 percent difference is over fitting; that means, 10 percent of over fitting. So, this R square is used for the linear model for the quadratic and the polynomial model, we used R square adjusted this is for linear model and this is for quadratic or polynomial.

So, these terms we need to calculate to see the closeness to see the fit of our model ANOVA tower the model we have generated ANOVA model would generate the regression models here and even we can have certain other advanced models for which we du conduct the ANOVA beforehand advance models might be new networks genetics algorithm artificial B column though all those techniques are kind of advanced techniques, but ANOVA can give the beforehand and the overall field that how would our model look like, how would over model behave what is the curve fitting in case of regression fitting we are talking about.

(Refer Slide Time: 12:12)

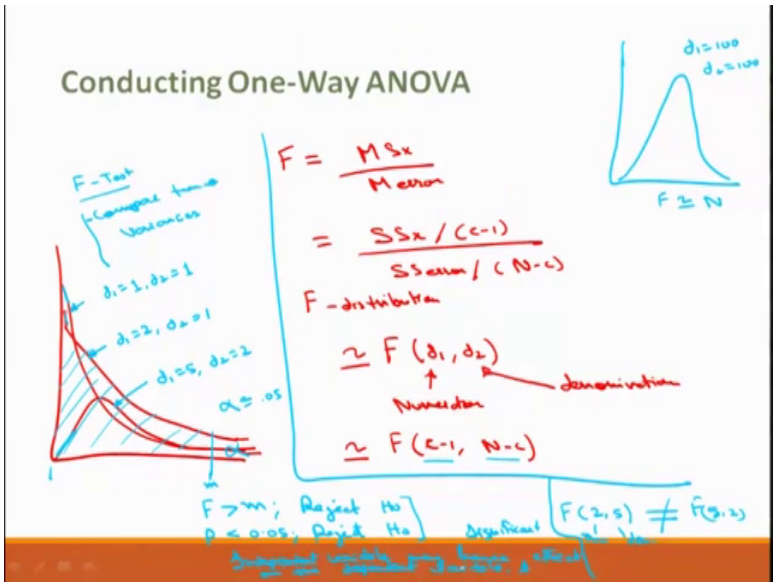


So, this is what ANOVA would help us. So, in one way analysis of variance, the interest lies in testing hypothesis that category measure all equal in the population and as I mentioned before; that means, all category means are equal; that means, the all come from the same population.

So, under this null hypothesis SS SS between and SS error come from the same source of variation, in other words the estimate of the population variance of Y this SY square root can be calculated as SS x by the degrees of freedom. So, this ss. So, this c minus 1 is the degrees of freedom this is also my degrees of freedom.

So, if you see this estimate of population variance this is nothing, but mean square due to x, which we denote as ms x and the second one is mean square due to error in which we have this error degrees of freedom and if we have the category or group degrees of freedom.

(Refer Slide Time: 13:38)



So, this is ms error; then we can have the F statistic value. Here the F which I just said is Ms x by Ms error this is equal to SS x by its degree of freedom and SS error by its degrees of freedom. So, this statistic follow F distribution. F distribution we write it like this F degrees of freedom 1 and degrees of freedom 2. Please make a note this d 1 is always numerator and this is denominator. So, in this case our degrees of freedom would be F C minus 1 and N minus C. So, F distribution is a writes queued distribution it is like this, something like this or may be like this or may be even like this, it depends upon the degrees of freedom for which we are trying to represent this F distribution.

Here we can have the alpha value that is, if F value if F statistic value is greater than this specific value then let me say this value is m. If F is greater than m we reject H naught and if it lies in this region this is excepted or in other words you can say if the probability this is p, alpha is my significance level alpha it may say if it is 0.05 for 95 percent confidence level or 5 percent significance level. If P value is less than 0.05, we reject H naught of the equal variances.

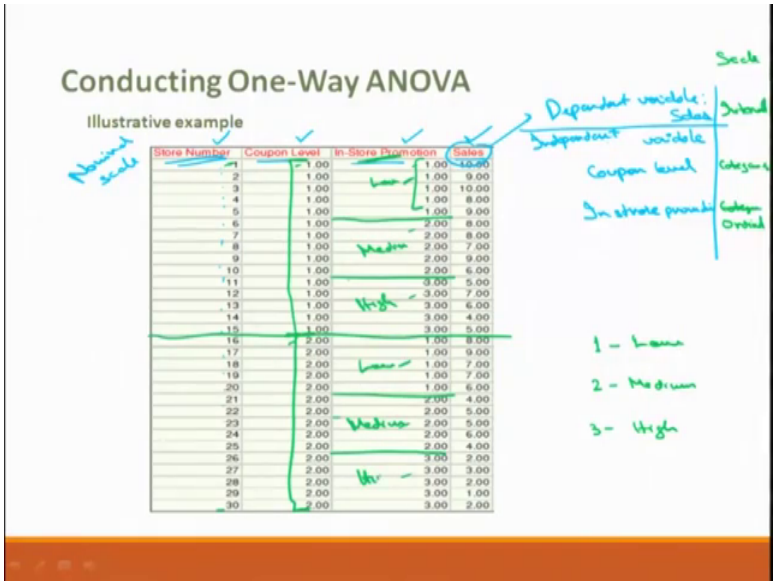
So, F test is the catch of all term for any test that it is F distribution. In most cases when people talk about F test they may are actually talking out the F test to compare two variances. So, F test is used to compare two variances and please note here that F let me say 2 5 is different this is numerator this is denominator this is not equal to F 52.

So, taking the freedom from numerator comes first for denominators comes second and why are these lines 3 lines 1 2 3 why are these different? This depend upon the kind of the degrees of freedom for the numerator and denominator like this line can be for d 1 is equal to 1, d 2 is equal to 1. The second line this which is having a less slope this can be for d 1 is equal to 2 and d 2 is equal to 1 the third line that is kind of a writes clued normal distribution, this third line is for t 1 is equal to 5 and t 2 is equal to 2.

So, for higher degrees of freedom we can even see the normal trend here, for let me say it would behave like this. If I say d 1 is equal to 100 and d 2 is equal to 100. So, very high degrees of freedom F is an approximation of normal. So, if this null hypothesis of equal category means is not rejected, then independent variables does not have significant effect on dependent variable.

This is rejected here; that means, the independent variable may have effect on the dependent variable. So, in this case the null hypothesis is rejected; that means, that the independent variable is significant, I can say here significant effect. It has significant effect on the dependent variable the comparison of category mean values will indicate the nature of effect of independent variable. So, let us try to compare category means here.

(Refer Slide Time: 19:55)



So, we have this data here, in which we have this store sales the in store promotion and coupon level and this is store number.

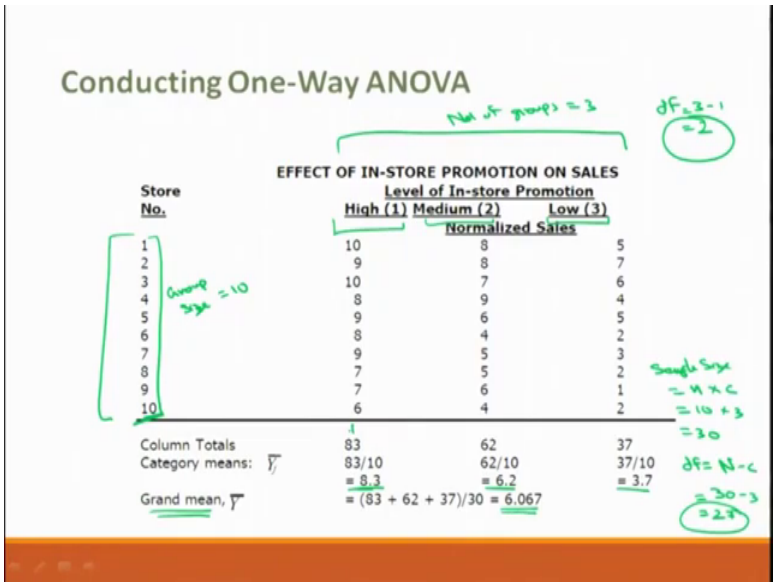
So, let me recall this test that we discussed. First we will try to identify the dependent and independent variable then will decompose the total variation total variation in to between groups and within groups, then we measure the effects that is eta square, omega square, R square then will test this significance for the specific level of significance for the specific significance will be set 5 percent significance level, then we interpret the results. In interpretation of the results we see that which is the most influential factor here. So, first let us try to identify, which is my dependent variable and what are my independent variables ok.

So, just please make a focused look on this, coupon level, in store promotion and sales, which is our dependent variable what is a store trying to look here. I think you are able to identify this the dependent variables would be the sales. So, dependent variable is sales independent variable. There are two independent variables, this store number is just a number of store you can call it as serial number or this is just the nominal data ok. Then independent variable is the coupon level and in store promotion and in store promotion.

So, if we see this we can put the types of this scales here, is it metric or non-metric scales dependent variables yeah the sales is a Indian metric scale, I will put the scale here it is interval scale and this one coupon level is nothing, but these are categories and in store promotion is also my categories. You can see 1 and 2 there are 2 categories here coupon level, and in store promotion they are 1 2 3 for coupon 1 and 1, 2, 3 for coupon 2, coupon 1 and this is for coupon 2 this is coupon 1 and 2.

So, if I say for in store promotion this 1 is low, 2 is medium and 3 is high. So, this is my ordinal scale here. So, I will better put it here this is the ordinal scale low medium and high level of promotion.

(Refer Slide Time: 23:49)

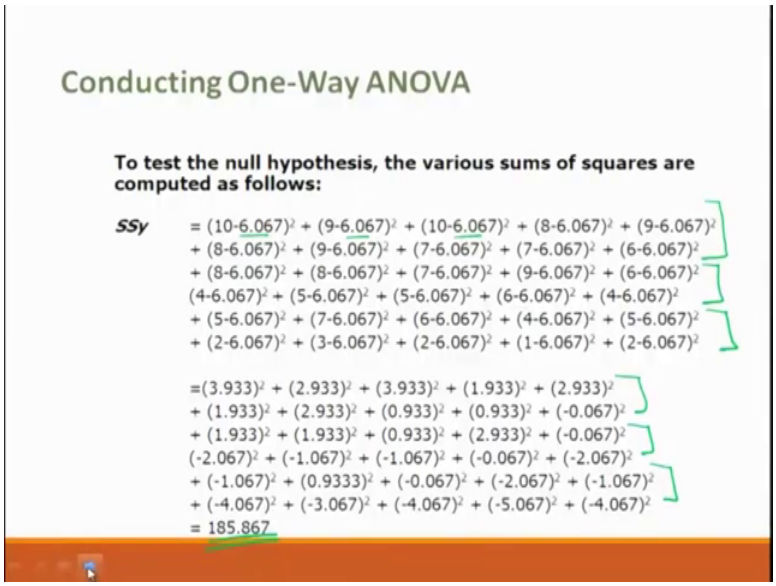


So, could I divide this in to these stores we have total 30 stores 1 2 30 stores? So, I have divided this is to 3 groups of in store promotion, that is low this is low and medium and number 3 this 3 three is high so; that means, we have got 3 categories here 3 categories high, medium and low and the category size is 10. So, this is my group size group size is 10 and this size that is number of groups this is equal to 3, this is equal to 10.

So, if I see the degrees of freedom for number of groups that would be the df would be 3 minus 1 is equal to 2 and we have sample size is equal to 10, that is group size N into C that is 10 into 3 is equal to 30 and degrees of freedom for this sample if you recall those were N minus C that would local to 30 minus 3 is equal to 3 minus 3 is equal to 27. So, we will use these 2 terms in the calculations.

So, this is the group mean. Group mean is 8.3, group total is 83 group mean for high, for medium, for low we have group means and we have grand mean of the whole data of the overall data this is the overall mean.

(Refer Slide Time: 26:12)

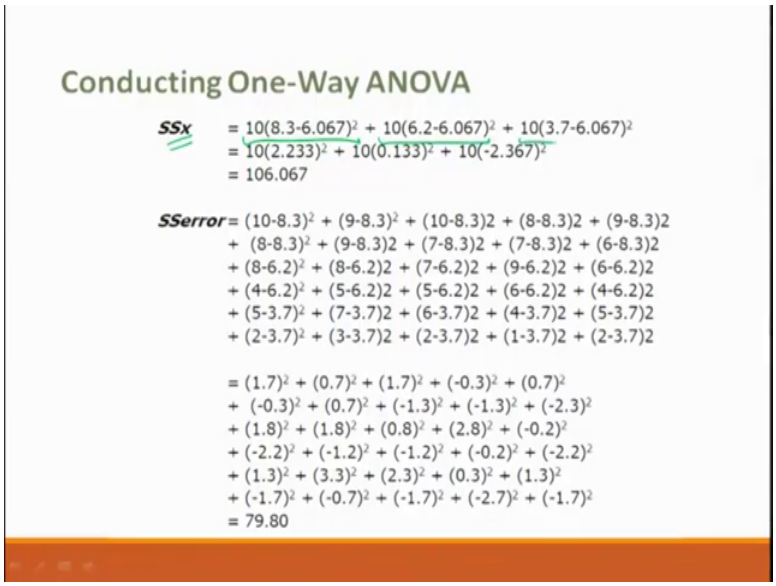


So, let us see the calculations SS total SS total SS y what was that? That was the variation of each value 10 minus the grand mean, this 9 minus grand mean 10 minus grand mean this 8 minus grand mean, 9 minus 10 grand mean so on we have 30 items and we keep on taking this squares of this, this is the sum of squares. So, this is for 10 minus the grand mean, 9 minus grand mean, 10 minus grand mean.

So, this is for group 1, group 2 and group 3 we got these values and we got the final total variation as 185.867. So, this is the calculation this is how the mechanism of ANOVA work. Actually when you just know this mechanism, when you work on the software you work on the packages like r package excel or if you have some spss or minute dial software's, you need not to do all this calculations this all would be done by the computers, but you should know that what is happening in the GUI what is the software doing what is the graphic use interface in that.

So, this thing would help you in that understanding how ANOVA works.

(Refer Slide Time: 27:42)

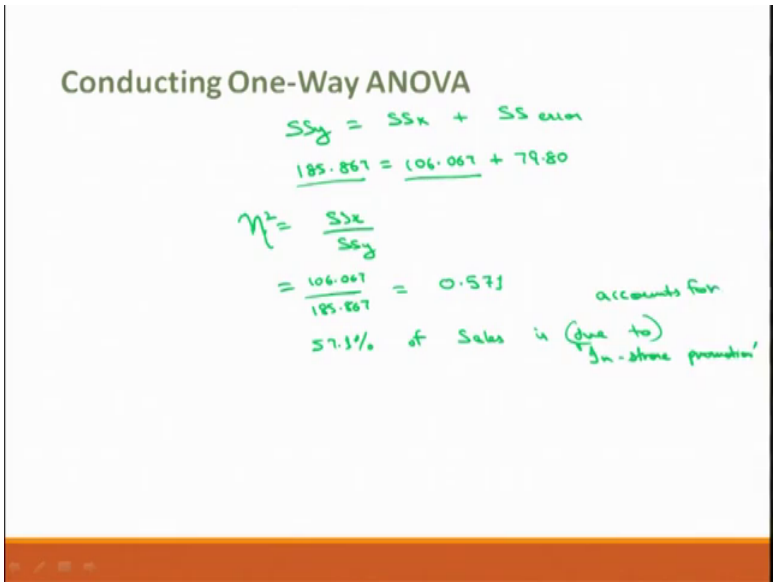


Similarly, we have SS between. SS between is the thing if you see this was the between means the grand mean and this means 10 minus 6.067, this 10 minus 6.067 or this is nothing, but difference in the categories and the grand mean; that is 8.3 this mean minus this mean. Then 6.2 minus 6.067 then 3.7 minus 6.067 whole square we have only 3 categories and we have total this is my group sizes 10 10 10.

So, this value comes down to 106.067. So, we have taken it for all the values you can see between. So, SS error. SS error is again the category mean here is for category one that is high level 8.3, category mean that difference of this mean from the elements 10 minus 8.3, this 9 minus 8 point 3, 10 minus 8.3 whole square this 8 minus 8.3 whole square so on we then move to second category 8 minus 6.2 this 8 minus 6.2 whole square plus 7 minus 6.2 whole square and so on then we move to category 3, similarly and we get the SS error and we got the total value as this much.

So, you might have noted that 106.067 plus 79.80 is equal to 185.867. So, that our decomposition of total variation that we saw before holds good here

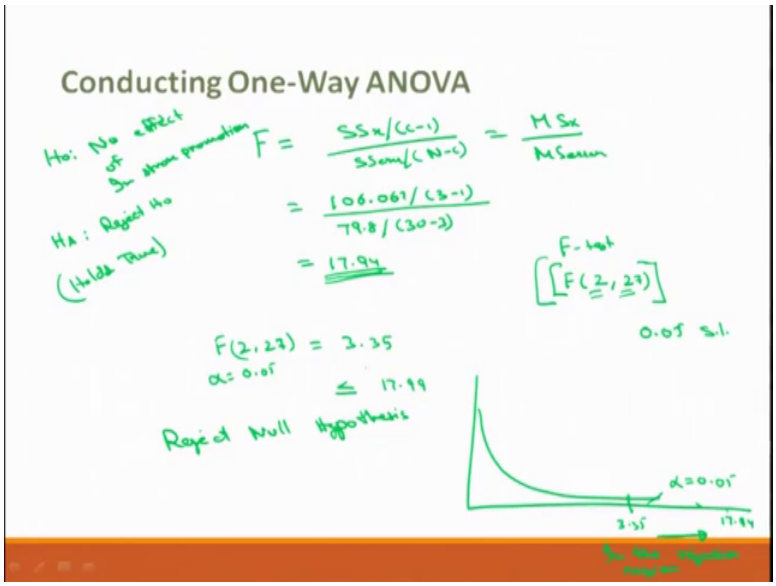
(Refer Slide Time: 30:00)



So, SS y is equal to SS x plus SS error so; that means, 185.867 is equal to 106.067 plus 79.80. So, the strength effects that is eta square that we can calculate by sd x by SS y this comes down to SS x value is 106.067 and SS y value is 185.867. So, this value is 0.571. So, in other words we can say that 55, 57.1 percent of variation in sales is counted by in store promotion here. 57.1 percent of sales is due to the variable this categorical variable in store promotion.

So, now, let us test the null hypothesis, we have got this statistic here and this is the value that test the null hypothesis using F statistic here.

(Refer Slide Time: 31:48)

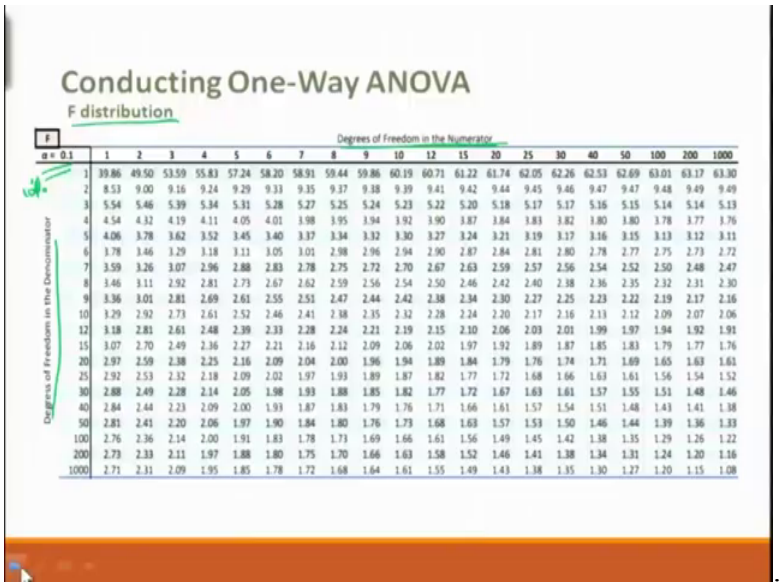


So, F statistic is SS x that is mean square for x that is SS x minus degrees of SS x over this degrees of freedom, then SS error by its degrees of freedom. So, this is I will put it again this is Msx and Ms error. This value comes down to SS x value is 106.067 SS error value is 79.8. 106.067 by 3 minus 1 and ms error is 79.8 by 30 minus 3.

So, this value comes down to 17.94. So, we can test this F statistic we can conduct the F test we have the F test using F distribution with F what is the degrees of freedom from numerator here? 3 minus 1 2 and degrees of freedom for denominator is 30 minus 3 27 for this we can see the value of F in the table and we will see whether this value is 17.94 is greater than the tab tabulated value here.

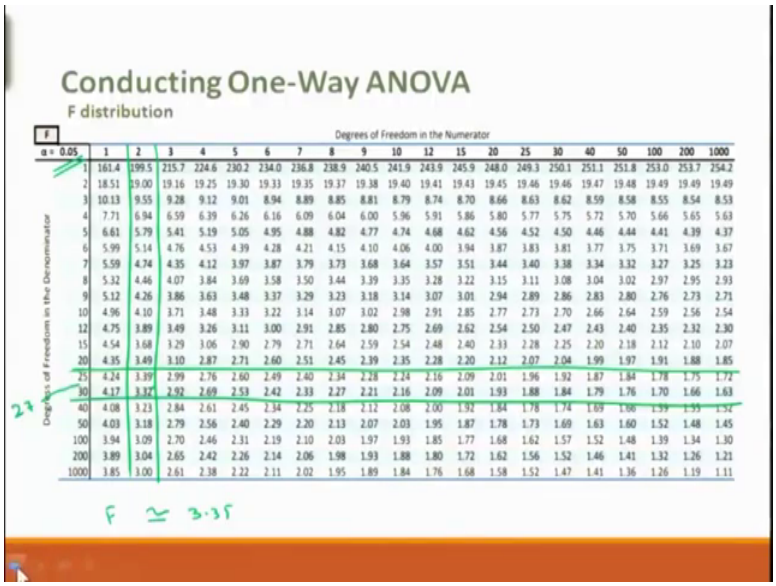
If it is greater than we will reject the null hypothesis, then we say we can say that yes whatever the hypothesis we are saying here this sale is due to I can put here better word is accounts for. So, let me test the significance. So, for 0.05 significance level we can see the F table here.

(Refer Slide Time: 34:08)



So, this is F distribution this is for specific alpha value, in this case alpha is 10 percent; 0.1 is 10 percent 0.1 significance level. So, in this case we have degrees of freedom in the numerator in the columns and degrees of freedom in the denominator in the rows, but this table is not for our test we are conducting because this is for specific alpha value. The second table yes this second table is for 0.05 that we have considered the alpha value here. So, for this value let me say two and 27 degrees of freedom.

(Refer Slide Time: 34:57)

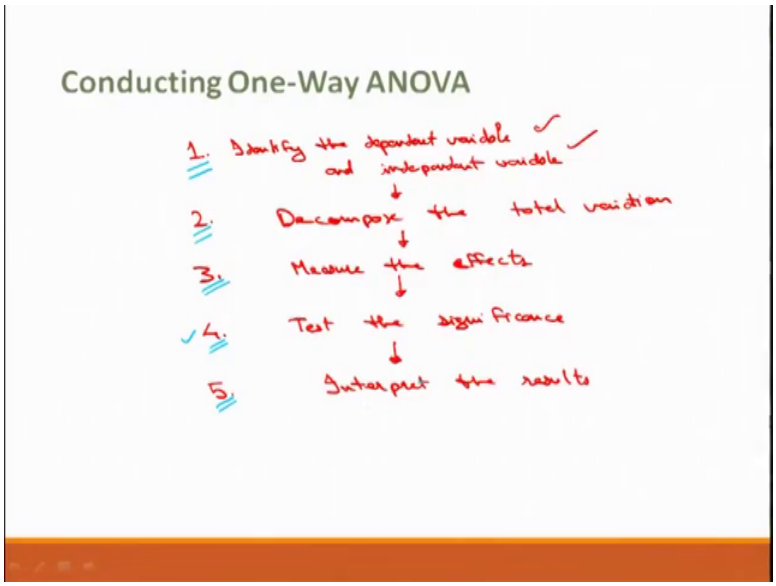


So, for two degrees of freedom we will select this column and for 27 degrees of freedom we have 25 and 30. We select these columns for 25 and 30 and we can extrapolate the value for 27 here. So, this value is 39 3.39 and 3.32 the in between value if we say that value comes down to F value, it is tabulated is 3.35. So, we found that the tabulated value that is F 2 27 value for 0.05 alpha is equal to 3.35.

So, then we see that this value is far less than 17.94. So, if I say this is my let me suppose my F distribution here would be this something like this and this is my alpha this is alpha is equal to 0.05 and this alpha value here is 3.35 and my F value the calculated F value is something very far that is 17.94 that is far from my significance level here. So, this is in the rejection region now we can conclude that we reject null hypothesis. Null hypothesis would be that no effect of in store promotion. So, this hypothesis is rejected and alternative hypothesis that is reject H naught.

So, this holds true; that means, we can say that the calculated value because is greater than the critical value the tabulated value therefore, the in store promotion has significant effect on these sales. So, this is the next strep we have tested in null hypothesis now we move to the final step that is interpretation of the results.

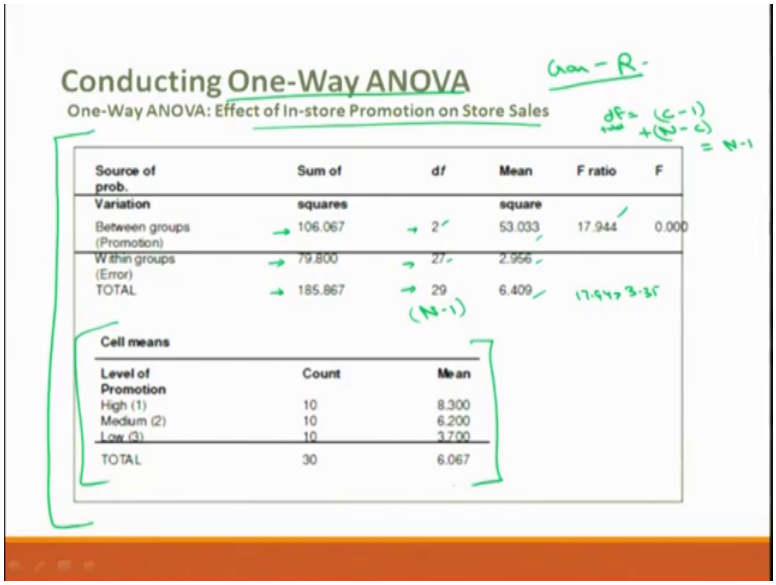
(Refer Slide Time: 38:08)



So, this is the significance level testing that is we rejected the null hypothesis and the interpretation of results is the same which I just said the in store promotion has significant effect on the sales here. So, we have completed the 5 steps identify, dependent

and independent variables, then we decomposed the total variation then we calculated the statistic value, then we test the null hypothesis. After null hypothesis tested we see that is the null hypothesis rejected or not. It is rejected, then we have interpretation of results that is our independent variable has significant effect on our dependent variable. So, this was our analysis of variance the mechanism here.

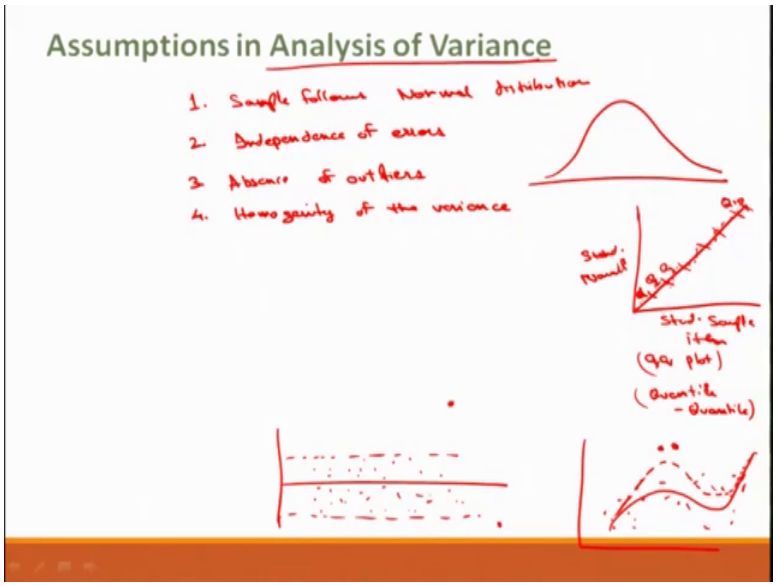
(Refer Slide Time: 39:04)



So, when you will do this in spss window or will do this in r package will show the examples of there as well we call it gran r you will see these kinds of windows. So, what do they say you just read the data from the excel or csv file, then we put the syntax for ANOVA and we get these values between groups within groups total variation degrees of freedom for these.

So, this is actually total degrees of freedom. Total degrees of freedom is nothing, but n minus 1 say it is actually this plus this is 2 plus 27. So, it is c minus 1 plus N minus C this is a category. So, if I plus this one degrees of freedom, total this comes down to n minus 1 then we have mean square values and we have F ratio. So, this was one way ANOVA. Effect of in store promotion on this store sales is significant because it tabulated on a critical value was 3 point 3 five. So, this 17.94 is greater than 3.35. So, this cell means also it is giving here.

(Refer Slide Time: 40:39)



So, there are certain assumptions in analysis of variance these assumptions are bound to be met if they are not met then analysis of variance that it does not hold good. So, what are assumptions? Number 1 sample follows normal distribution that is if I plot all the sample items, it should follow this normal distribution how do we test this? We actually do the quartile test we put the standardized normal distribution here and we have standardized sample items. And we see that what is the correlation between them, do the line is close to this 45 degree line if it lies like this one? Yes we can say that they are very much following the normal distribution.

So, this is known as qq plot, qq is nothing, but the quantile quantile plot. This note I have put quantile quantile here because we have we can divided into number of parts here. If we say the quartiles like in the box plot we have 3 quartiles q 1, q 2 and q 3 we have divided the whole data into four components here and the q 2 is was the median if you remember.

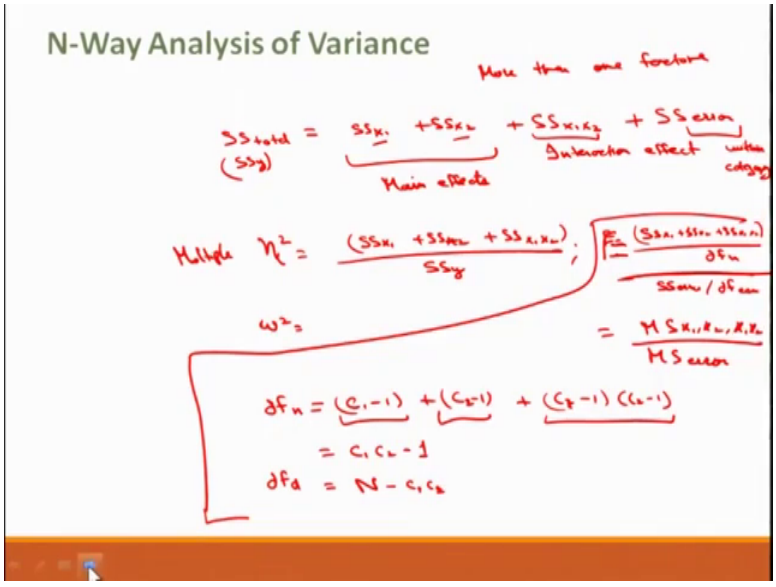
So, in this case we quantile quantile plot we can say each let me say if they are 10 different parts here this is q 1 to q 10 and equal number of data points are here in each slot here q 2 q 3. So, this sample follows normal distribution we will see this graphical this visualization, when will see the assumptions of ANOVA in r package we will see r is the assumptions being met or not. So, next is independence of errors that says that the errors between the cases are independent of one another there is no dependence of errors

the errors are totally independent and there is some random chance that errors are coming.

So, next assumption is that absence of outliers; absence of outliers. For instance if this is my data and this is the total variance plot here this could be the outlier these points might the outlier; So, absence of outlier because this outlier would affect my data. So, for example, if this is my model, this is the data, and this is the model here. So, if I have a few outliers here 1 or 2 points which are far behind the model would come like this. Like these outliers would affect the model it will try to attract the model the line to r itself this outliers. So, that would affect now this point I was mentioning here this is the variance here.

So, it states that the variance has to be homogenous number four is homogeneity of the variance that is population variance in different levels of each independent variable are equal. So, they are equally distributed here.

(Refer Slide Time: 44:57)



So, next comes the n way analysis of variance; n way analysis of variance is nothing, but we have more than ONE factor like we had only ONE factor that is in store promotion in the previous example. in n way analysis of variance it is often concerned with the effect of more than ONE factor simultaneously for example, the advertising levels and price levels both how does it depend upon the brand sales for example, they are the we divide the advertisement the level of advertisement into 3 categories high medium low, it

divides the price of product in to 3 categories ONE high medium and low and there is a 2 categories here, that is a 2 factors we can say and what is the effect of these two factors on this sales.

In this case 2 way analysis of variance would happen another example I can take here I can extend the example which I took before the familiarity with the NPTEL courses and NPTEL portal and number of students who register if I say this was the 1 factor familiarity high medium low was 1 factor the number of courses taken the number of students in the courses is the measure dependent variable here; If I take another factor that familiarity with the NPTEL courses second factor if I take reputation of this specific institution specific IITs. IIT Kanpur, IIT Madras, IIT Kharagpur, IIT Delhi these are 2 factors reputation, high and low depending upon the quality of the courses which are delivered by the specific institutions.

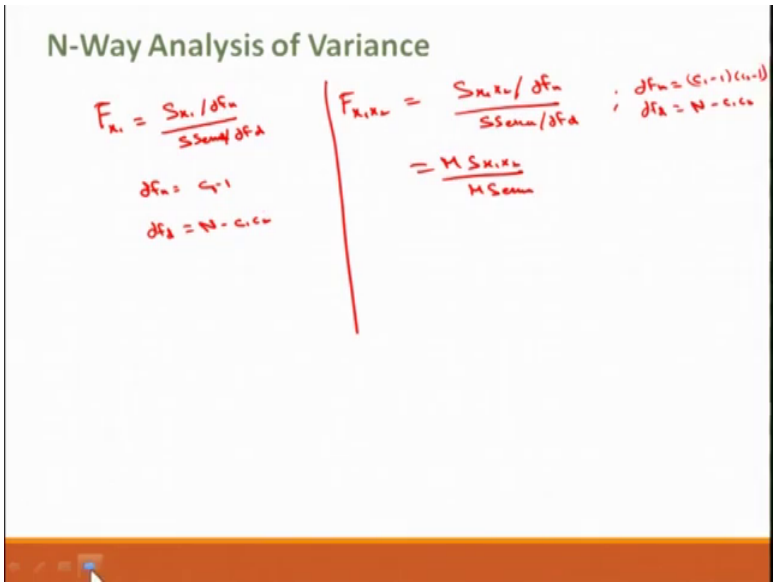
So, we have the quality of the courses depending upon the institution and the familiarity with the NPTEL portal, the NPTEL courses are there. These two factors if are there in again we have 2 way ANOVA. So, I will try to continue my previous example here and would calculate the SS total as SS due to x 1 and SS due to x 2 and SS due to x 1 and x 2 combined this is known as interaction effect, these are known as main effects.

So, we had 2 factors here coupon level and in store promotion. We did first in store promotion and sales. Now will do the in store promotion and coupon level on sales they have coupon level 1 that 2, 2 levels of categories here they have 3 levels of categories here in store promotion, how does this effect our sale we will see this. So, not to 4, but here this is actually the perfect model if we do not have the error here we need to put SS error, that is SS within category.

So, in this case we have the overall effect that is known as multiple eta square then we have omega square as well here that is the individual effects here. So, this eta square for if I said 2 ANOVA x 1 and x 2 here, this eta square is calculated as SS x 1 plus SS x 2 plus SS x 1, x 2 by SS total or SS y. And in this case F statistic would be SS x 1 plus SS x 2 plus SS x 1, x 2 by the degrees of freedom for n this is F statistic here over this is mean square for x 1 x 2 and x 1 x 2 combined this over the SS error by degrees of freedom for error. So, this is nothing, but mean square for x 1 comma x 2 comma x 1 x 2 by mean square for error.

So, degrees of freedom for numerator here would be, the degrees of freedom for category 1 plus degrees of freedom or category 2 plus combined degrees of freedom for category 1 and category 2 category 1 category 2 combined category 1 category 2 that comes down to C 1, C 2 minus 1. And degrees of freedom for denominator for this F test would be degrees of freedom for total all the values minus C 1, C 2 ok. So, we can calculate this F statistic to see the overall effect we can even see the F value for the individual effects.

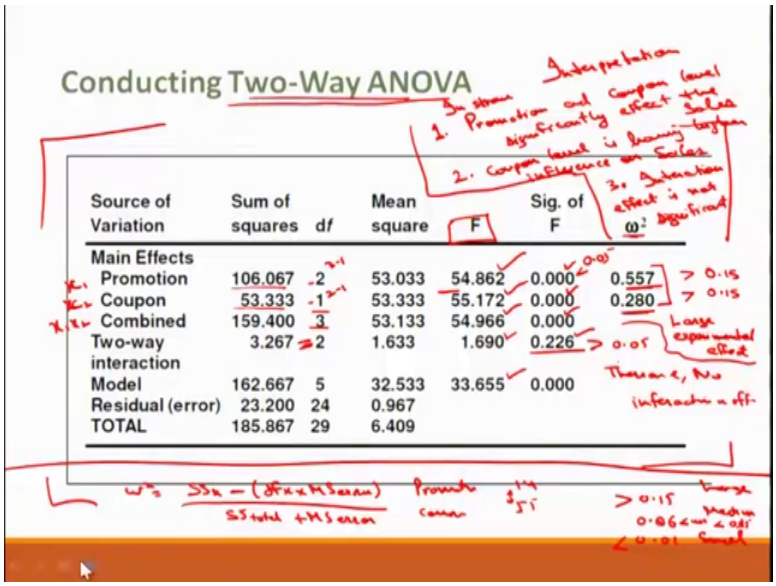
(Refer Slide Time: 50:40)



For example interaction effect F 1 and x 1 x 2 that would be S for x 1 x 2 by degrees of freedom for a numerator here by SS error by degrees of freedom for denominator. So, this is MS x 1 x 2 by MS error. So, for this is this was for the interaction effect for the main effect let me say for x 1, this would be again Sx 1 by degrees of freedom for the specific x 1 and SS error by degrees of freedom for the denominator.

In this case the degrees of freedom for x 1 x 2 is degrees of freedom for numerator here is C 1 minus 1 into C 2 minus 1 and degrees of freedom for denominator here is N minus C 1 C 2. And in this case degrees of freedom for numerator would be for this category 1 C 1 minus 1 and degrees of freedom for denominator here would be N minus that is error N minus C 1 C 2.

(Refer Slide Time: 52:14)



So, if I consider the two categories here, this kind of table you can do the all the calculations if you want you can do that in excel or may be manually you can do, but I will show you the table that is an SPSS output here. So, we have the promotion the coupon and the combined this is category 1 this is category 2 and this is actually x 1 x 2 and this is x 1 x 2 combined.

So, we had for promotion this is the sum of squares and for coupon level if you see this is the sum of square we had three categories this is 3 minus 1 is equal to 2, we had two categories 2 minus 1 is equal to 1, total is 2 plus 1 is 3 and combined is nothing, but sum of this, this is sum of these 2 interaction we have c this in to this 2 in to 1 is equal to this 2, 2 into 1 is equal to 2.

So, what we can see we have the F values for 2 factors and combined F value and we also have the 2 way interaction and if total for the model here. So, we can see here that the F value as in the significance of F the probability of F is less than 0 less than 0.05 here this one, this one and this one. These two that is the promotion and coupon level both are significant the interpretation would be number 1 promotion it is the in store promotion and coupon level significantly affect the sales this is number 1.

Number 2 we can see that which has more significant, which has more influence on this in that case again F is the criteria that would decide. In this case coupon level is having lager value, but these 2 values are close 54 and 55 are close, we can say that coupon

level is having higher influence on sales. If I suppose that this for promotion, they are value let me say it would have been some value like 14, couple level it would identify.

So, we can check this significance also is the difference between these 2 values significant or not. Then that third interpretation here is 1 2 and 3; third interpretation here is that the interaction this is the interaction, this value is greater than 0.05 therefore, no interaction effect. So, the third interpretation I will put here, interaction effect is not significant.

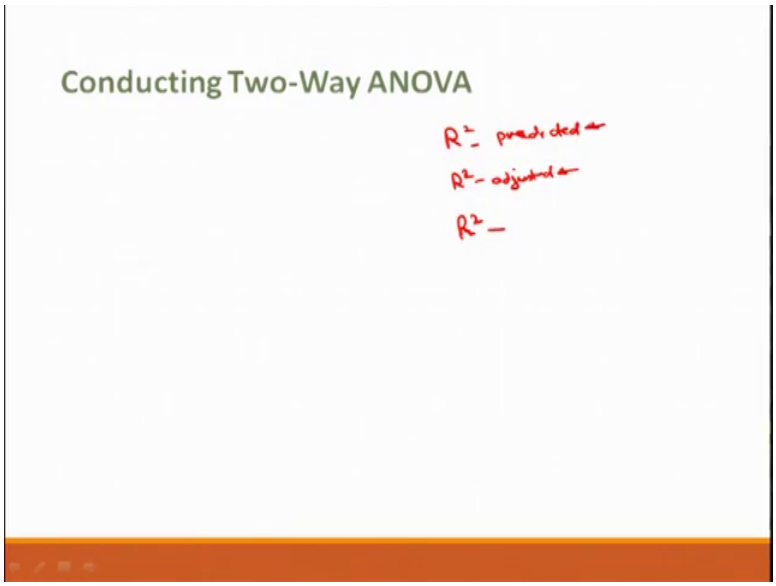
So, these are the interpretations or the results of our ANOVA table also we have omega square here. Now what is omega square here? Omega square indicates what portion of variation in the dependent variable is related to a particular independent variable, this has the proportion ok. So, this is calculated only for this significant variables here, because this is less than 0.05. So, this is calculated omega is calculated omega square is equal to the SS x minus degrees of freedom for x in to MS error by SS total plus MS error.

So, if in this case if I calculate the value of omega square for promotion and for coupon the values comes from to 0.557 and 0.280 respectively. So, as a guide to interpret this term, we can say that in large experiment effect produces index of 0.15 or greater. And a medium effect produces an index around 0.06 and a small effect produces an index of 0.01.

So, this is large, this is medium and this is small, we are talking about the omega square here ok. If it is less than 0.01 it is small, if it is between 0.06 is less than omega square is less than 0.15. So, in the present example 0.557 and 0.280 are both greater than 0.15. Therefore, it has large effect, large better to put large experimental effect. So, this was analysis of variance for 2 factors or we call it 2 way analysis of variance. Similarly we can conduct n way analysis of variance and multiple factors could be identified we can have the list of factors, but please mind it.

The more we add, the more factors we have more is the value of R square that is inflated. So, value of R square would be higher four number of factors, in that case R square adjusted value has to be seen to see the effect of the model and also we have R square predicted, and also we have R square we have R square predicted, R square adjusted and the normal r square.

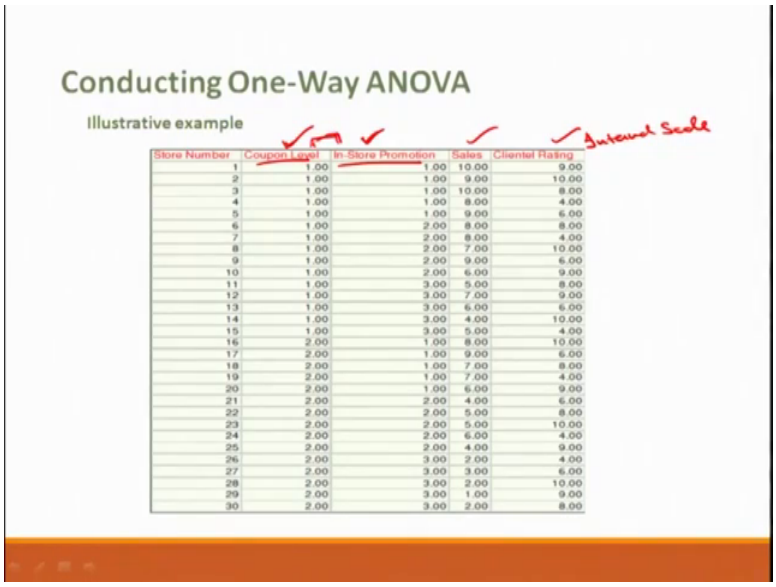
(Refer Slide Time: 59:22)



So, this R square looks good but they can be serious problem with an over fit of the model. For one thing the regression coefficients represent the noise that rather than the general relationship of the population, that is the un accounted factors could effect this one.

So, this adjusted R square can compensate that. More over the predicted R square value is designed to detect the over fitting as well. So, next thing is analysis of covariance.

(Refer Slide Time: 60:06)



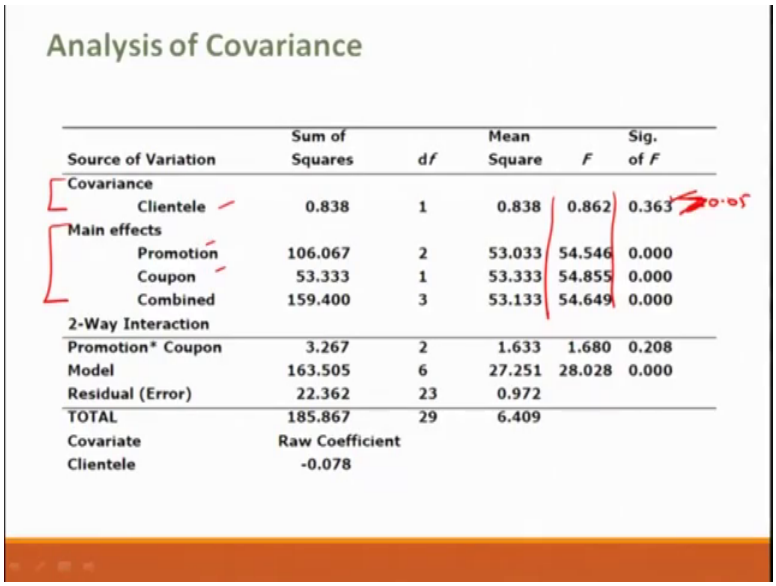
Store Number	Coupon Level	In-Store Promotion	Sales	Clientel Rating
1	1.00	1.00	10.00	9.00
2	1.00	1.00	9.00	10.00
3	1.00	1.00	10.00	8.00
4	1.00	1.00	8.00	4.00
5	1.00	1.00	9.00	6.00
6	1.00	2.00	8.00	8.00
7	1.00	2.00	8.00	4.00
8	1.00	2.00	7.00	10.00
9	1.00	2.00	9.00	6.00
10	1.00	2.00	6.00	9.00
11	1.00	3.00	5.00	8.00
12	1.00	3.00	7.00	9.00
13	1.00	3.00	6.00	6.00
14	1.00	3.00	4.00	10.00
15	1.00	3.00	5.00	4.00
16	2.00	1.00	8.00	10.00
17	2.00	1.00	9.00	6.00
18	2.00	1.00	7.00	8.00
19	2.00	1.00	7.00	4.00
20	2.00	1.00	6.00	9.00
21	2.00	2.00	4.00	6.00
22	2.00	2.00	5.00	8.00
23	2.00	2.00	5.00	10.00
24	2.00	2.00	6.00	4.00
25	2.00	2.00	4.00	9.00
26	2.00	3.00	2.00	4.00
27	2.00	3.00	3.00	6.00
28	2.00	3.00	2.00	10.00
29	2.00	3.00	1.00	9.00
30	2.00	3.00	2.00	8.00

So, let us see this example unit we have this two factors coupon level and in store promotion these were the factors categories, and there is another variable that is cliental rating what is the rating of the client and this is in interval scale. And then as I mentioned earlier if we have the interval scale as well as categorical scale, the analysis of variance is then called as analysis of covariance.

So, when examining the mean values of the dependent variable related to the effect of the controlled independent variable, it is often necessary to take into an account the influence of the uncontrolled independence variable. For example, in determining how different groups exposed to different commercials evaluative brands it may be necessary to control the prior knowledge of the customer to of the groups here.

In determining how the price level will effect, a house hold consumption it may be sensual to take the how house hold size into account. For instance in this case the cliental rating if the client is having high rating, the purchasing by a this client might be higher this might effect. So, in this case we conduct analysis of co variance.

(Refer Slide Time: 61:27)

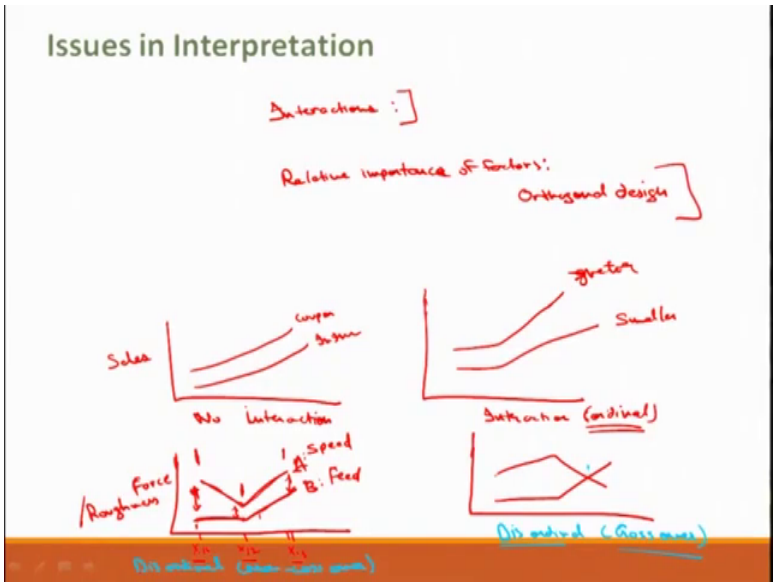


So, this is the table here, these are the main effects here, this is the covariance, this is actually the interval scale. So, these are in my categorical scale the factor factors here.

So, in this case we have again the F statistic value. So, we can do the similar interpretations from this table as well. In this case the cliental rating is not significant,

this value is greater than 0.05; that means, that cliental rating does not have or does not influence my sales here. So, next there are certain issues interpretation, there are interactions and there are relative importance of the factors.

(Refer Slide Time: 62:18)



First I will take the relative importance of factors experimental design are usually balanced. In that each cell contains the same number of respondents, as I had 10 stores in each cell here, but sometimes this is not possible. So, this kind of design when each cell is having same number of respondents, that is known as orthogonal design. So, it is possible to determine the relative importance of each factor in explaining the variation independent variable in this case.

So, interactions are when the 2 variables are interacting each other. So, for instant if I plot this here, let me say sales and the 2 variables here my coupon and in store promotion, this is not interacting with each other; like if with increase in this also increasing there is no interaction.

We can have some interaction in certain cases for instance, if it is like this one. In this case we have a interaction that is ordinal, it has this one greater this one is smaller. That is the lines are not exactly parallel here. But we can see that is the interactions significant or not. All though interaction is there, but significance also needs to be tested, then we have dis ordinal interaction also. This is the interaction for instance if its like this one. See if this is increasing if I say they are 3 values x 1 1, x 1 2 and x 1 3 this 1 and I could

put A and B, this is increasing at this point and again increasing the B. B is increasing and again the slope is higher and it is increasing A is first decreasing and then increasing. So, there is a difference.

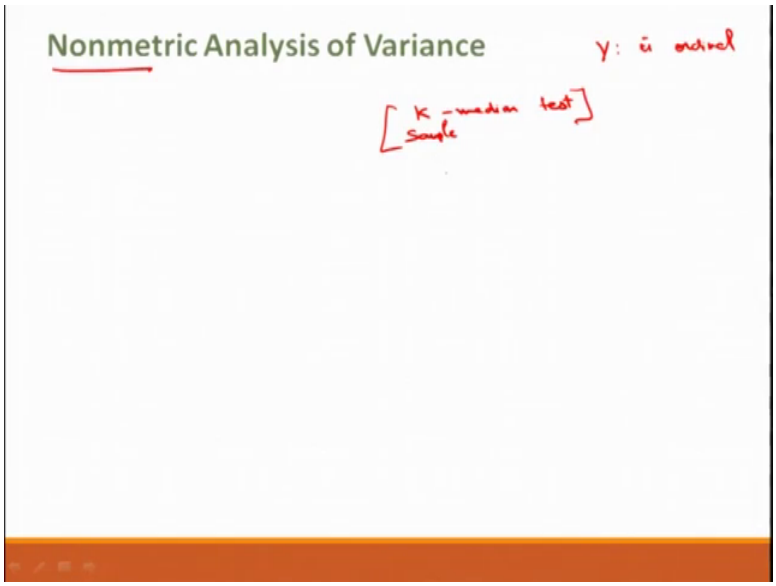
So, this is large, this distance is large, this distance is small again this distance is large. So, there is an interaction. So, these things for example, if I do machining, if you know the mechanical people do machining they cut the material. They change the tool cutting speed, they change the tool speed cutting speed is the for instance the rotation of the work piece, which is being cut here it is if the work piece is rotating it is cut by the tool.

The feed is the rate of cutting here of this tool this can have a transient behavior for instance a certain at lower speeds at lower speeds, the cutting; that means, am trying to see the cutting force here or may be roughness of the work piece and you response I can choose here, speed feed also I have put A as speed and B as feed.

So, here I can see the feed is increasing from level x 1 1 to level x 1 2, but increase from level x 1 2 and x 1 3 is higher this slope is higher; however, at lower speeds this is low speed this is high speed and lower speed the force is higher, at intermediate speed force is lower and at high speed the force is again higher. So, this has an interaction the previous. So, the previous one was ordinal.

So, this is dis ordinal dis ordinal, but non cross over. Sometimes we can have the case when if I say this same case for different machines or for different material for different tool combination tool work combination or for different environment, it can behave something like this as well. So, this is again dis ordinal, but we have cross over. So, this is dis ordinal and also we have cross over here, it is crossing. So, this transient behaviors can be like this as well.

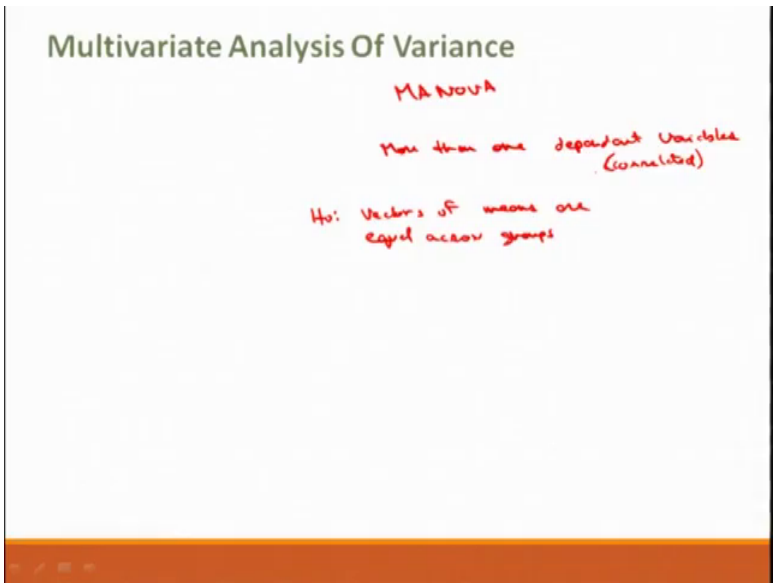
(Refer Slide Time: 67:58)



The for the non-metric analysis of variance there are certain tests like K median test, this is actually known as K sample median test. So, this is non metric analysis of variance this examiner difference in the central tendencies of more than 2 groups, when the dependent variable is measured on an ordinal scale in this case Y or dependent variable is ordinal. So, it is non metric.

So, in this cases we can have these kinds of test, which are I can say that out of the scope of this course that we can provide notes to you for this.

(Refer Slide Time: 68:47)



Also we have multivariate analysis of variance that is MANOVA; so, what if MANOVA multivariate analysis of variance is similar to ANOVA, except that we have more than one dependent variable. So, in MANOVA the null hypothesis this is that the vectors of means of people depend with peoples are equal across groups; vectors of means are equal across groups. So, multivariate analysis of variance is appropriate when there are two or more dependent variables that are correlated.

So, next we will meet in the r session for analysis of variance we will take the certain examples, here we will do ANOVA test for 2 factors for 1 factor for 2 factors and for multiple factors and we will see how to interpret results, we will see how the assumptions are being met and what to do if assumptions are not being met here.

Thank you.