

**EXAMINATION OF DIFFERENT**

**STATISTICAL METHODOLOGIES**

**Submitted By**

Sai Chethan Singu – x18181937

Data Analytics – JAN/Batch B – 2020/2021

**(1) Binary Logistic Regression With PCA**

Logistic regression is a supervised classification algorithm which is used to predict the probability of dependant or target variable. When there is only two possible categories of dependent variable (Yes or No), then it is classified as Binomial logistic regression and more than two categories of dependent variables is known as Multinomial logistic regression. In this model principle component analysis is done prior to the logistic regression to increase the probability of prediction and to reduce the number of dependent variables.

**DATA SET USED:**

The various factors that affect the job opportunities that includes immigrants, diversity in workplace, outsourcing, automation, using foreign products, and increased exports have been extracted from the dataset for the analysis. After the removal of null values, resultant dataset consists of 108 records.

<https://www.pewresearch.org/global/datasets/>

* Job opportunities

Dependent Variable

* Immigrants
* Diversity
* Outsourcing
* Automation
* Exports
* Foreign Products
* Women in workforce

Independent Variables

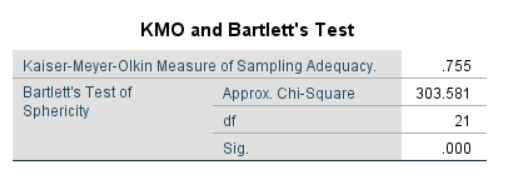
**OBJECTIVE OF THE ANALYSIS:**

The objective is to carry out the Principal component analysis and then logistic regression to predict the dependent variable (Job opportunities) using independent variables that influences the most for US. By performing PCA, the number of factors can be reduced to smaller number and thus performing logistic regression to categorize the job opportunities.

**PRINCIPLE COMPONENT ANALYSIS:**

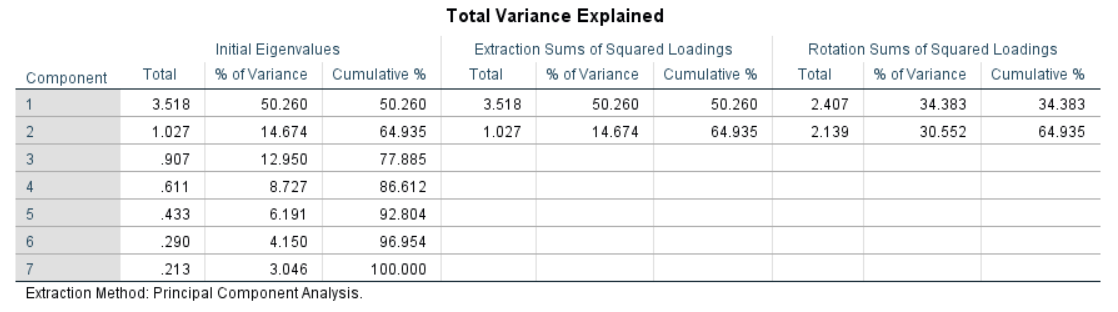
**ASSUMPTIONS FOR PRINCIPAL COMPONENT ANALYSIS:**

**Step 1:**

****

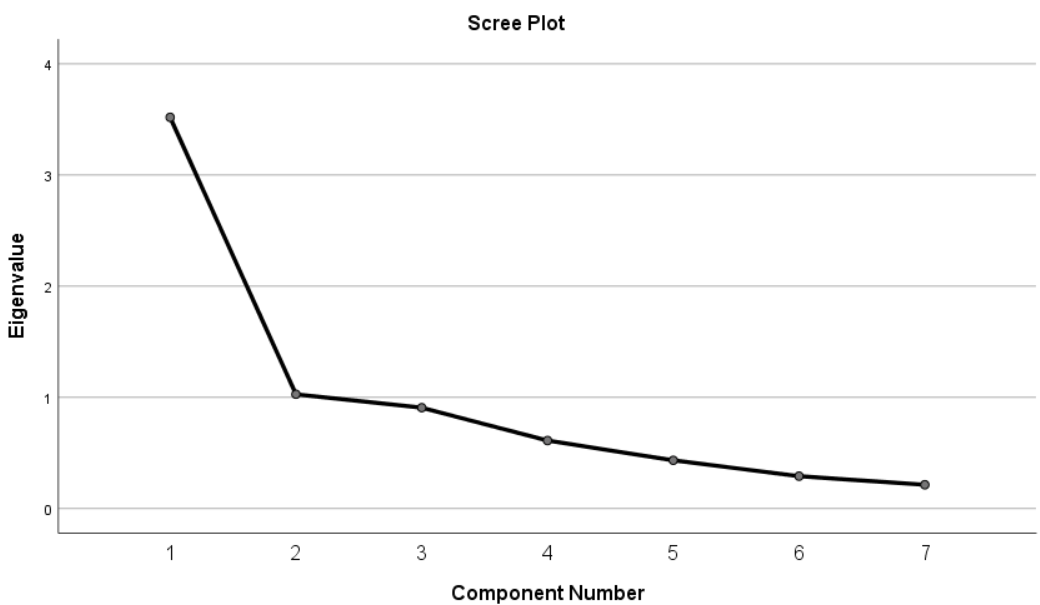
The primary step to verify the data is suitable for performing factor analysis is to check the KMO value. In this data set Value is 0.755 which is greater than 0.6. Bartlett’s test is also significant (.000). This shows that factor analysis is suitable for this data.

**Step 2:**

****

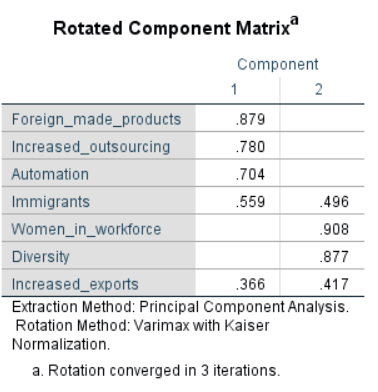
By seeing at total variance explained table, the cumulative percentage of variance is 64.93 that is total variance by two components. The extraction sum of squares shows that dependent variables like immigrants and women in workforce are the most influential components among others.

**Step 3:**

****

By looking at the shape of the scree plot first and second components have explained much variance when compared to the other five components.

**Step 4:**

****

Component matrix is rotated by principal axis using Varimax rotation method. From the table foreign made products, outsourcing, automation comes under component one and the remaining comes under component two. Immigrants and increased exports are cross loaded in both the components. By performing principle component analysis, only two components have been derived for the purpose of meaningful interpretation. The two regression factors obtained in principle component analysis are taken as independent variables for performing binary logistic regression.

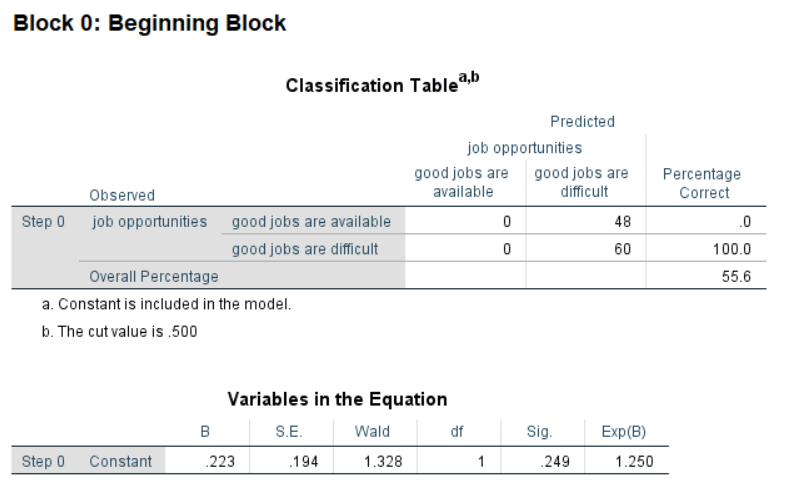
**BINARY LOGISTIC REGRESSION:**

**SAMPLE SIZE:**

The sample size is calculated by the formula N > 50+8M where ‘M’ Is the number of independent variables in the dataset. In this dataset the sample size is 108, which satisfies the assumption.

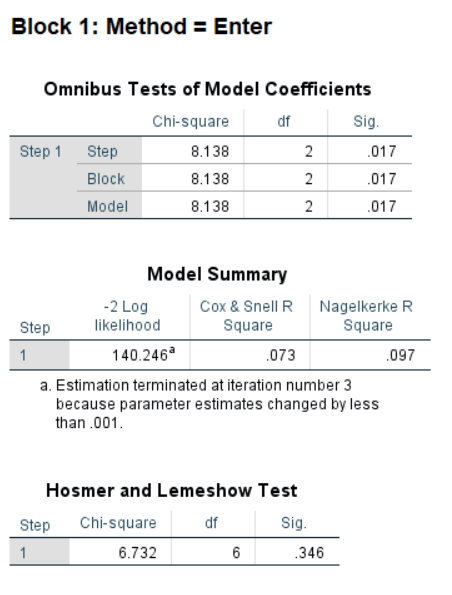
**ANALYSIS OF THE BINARY LOGISTIC REGRESSION MODEL:**

**BLOCK 0:**

****

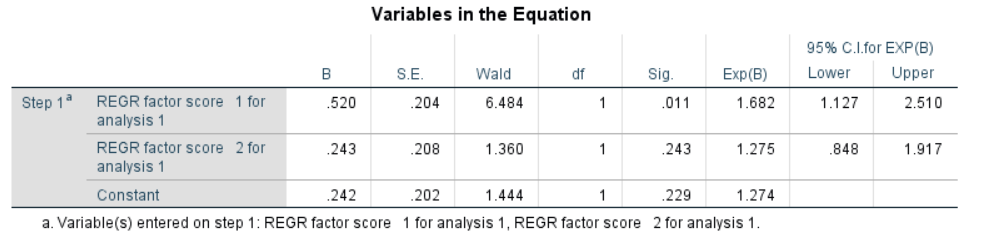
Block 0 predicts the results of the analysis without using any of the independent variables. From the classification table above, the model predicts with accuracy of 55%. B value is useful for calculating the probability of the cases that falls under which category. In this model, the positive B value indicates that there will be increase in the probability of dependent variable with increase in independent variable. The Exp(B) denoted odds ratios for each independent variable which is 1.25.

**BLOCK 1:**



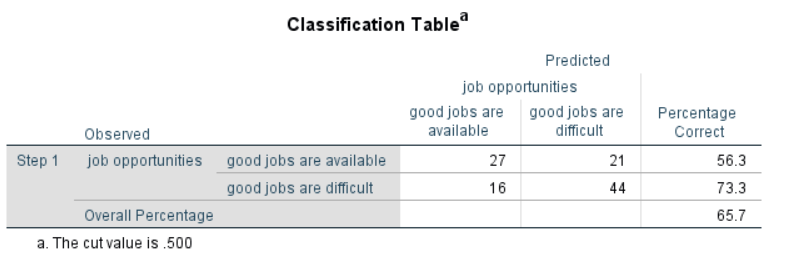
In block 1, predictor variables are tested. The Omnibus tests of model coefficients signifies the performance of the model. Since it is less than 0.05, it is highly significant. In model summary table, Cox’s and Snell R square and Nagelkerke R square indicates the range of variance by the model in dependent variables. In this model, it suggests the variance ranging from 7% to 9% of total. In Hosmer and Lemeshow. test, Chi-square and significance values are taken into account to test the model fitness in which significance value should be greater than 0.5. In this test Chi-square value of 6.73 and significance value of 0.34 indicates that, it supports the model.

**VARIABLES IN THE EQUATION:**



This table tells the importance of each independent variables. The B value indicates the change in dependent variable with respect to independent variable. Regressor factor 1 is statistically significant in predicting the job opportunities by having significance value less than .05 and the other variable is not significantly contributing. The Exp(B) column values are the odds ratios that means odds of having presence in any category when independent variable is increased by one unit. In our model, the odds of people to have good job opportunities is 1.68 times greater. The last two columns represents 95% confidence intervals.

**CLASSIFICATION TABLE:**

****

This table tells how model predicted the desired categories that job opportunities are available or difficult. As a conclusion out of all, people who feel good jobs are available are 27 and people who feel good jobs are difficult to get are 44. The other two corresponding values are wrong predictions by the model. The model correctly predicted 65.7 percent when compared to 55.6 percent in Block 0. The sensitivity (True positive rate) of this model is 73.3 percent and the Specificity (True negative rate) is 56.3 percent.

**CONCLUSION:**

The availability of job opportunities is predicted using binomial logistic regression performed in SPSS software. The variance of the model is explained with Cox and Snell R square and Nagelkarke R squares are 7 and 9 percentage respectively. B coefficients and exp(B) are within the acceptable range. Hosmer and Lemeshow values are also significant and within the acceptable range. The model predicts the job opportunities for good jobs with 65.7% accuracy.

**(2) ANOVA**

Analysis of variance is performed to examine the differences between the mean values of dependent variables. It is the total sum of squares of squares between and within the groups. In this model one way analysis is performed.

**DATA SET USED:**

The dataset contains the information of Intellectual property rights and its licensing. Trademarks that are owned by different companies in Europe data was taken along with number of employees as size. The dependent variable (continuous) is number of trademarks and independent variable is categorized by number of employees. Company having employees ranging from 10 to 49 are categorised as 1, employees ranging from 50 to 249 as 2 and employees ranging from 250 and above as 3. A dummy variable has been created next to the independent variable with renamed categorical values.

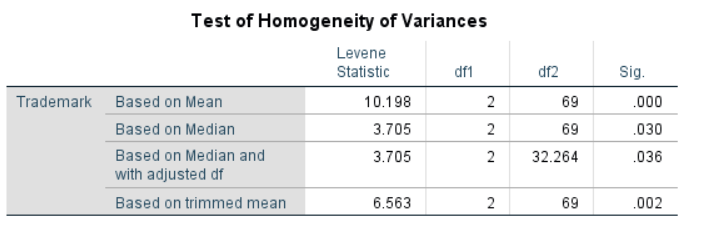
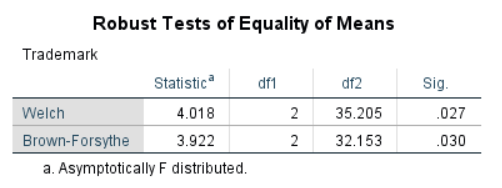
<https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=inn_cis10_ipr&lang=en>

**OBJECTIVE OF THE ANALYSIS:**

Is there any significant difference associated in the number of trademarks owned with respect to number of employees in the company.

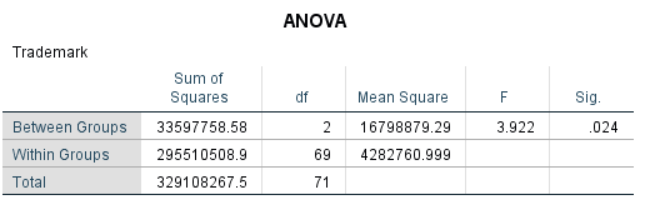
**ASSUMPTIONS:**

**1:**

**** ****

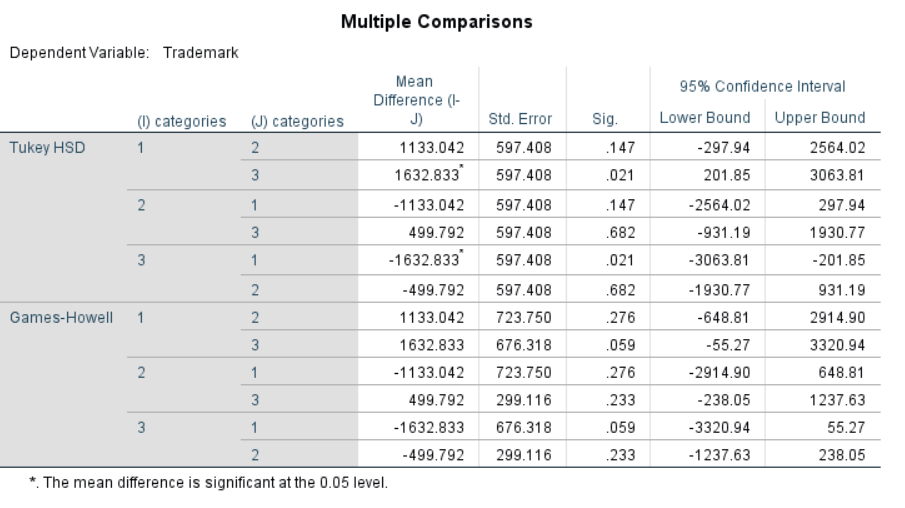
As the significance value is not greater than 0.05, Levene’s test of homogeneity assumption is not satisfied. As a result of that Robust tests table is taken into consideration where it shows the value .024 which is satistically significant.

**2:**

****

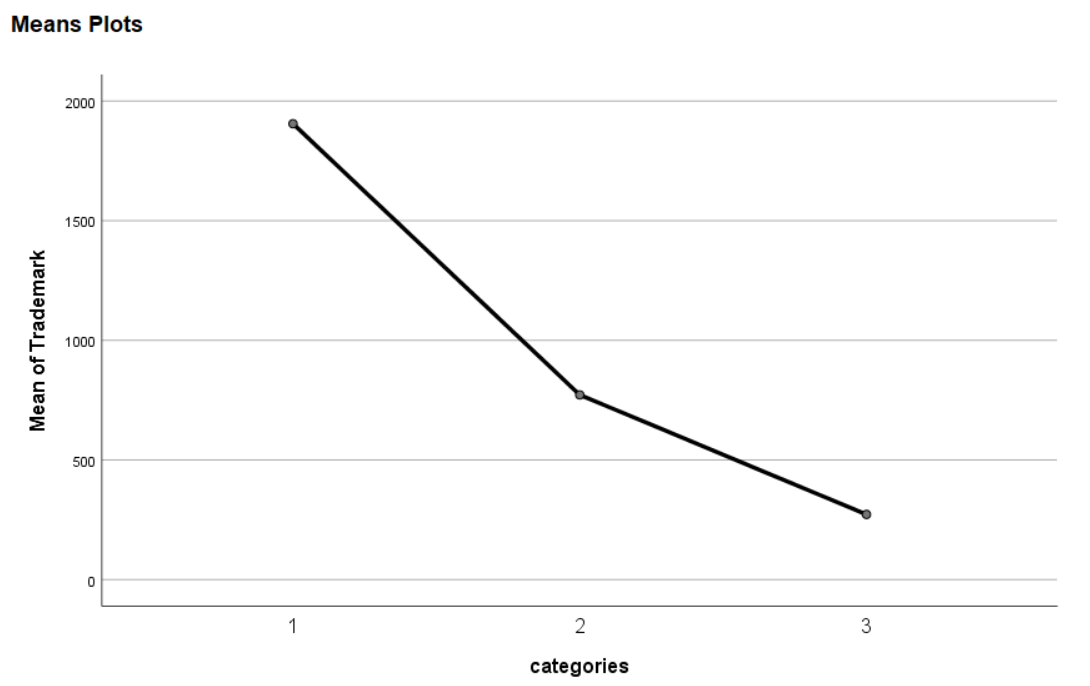
From Anova table, the significance value is .024 which is less than .05. This indicates that there is significant difference among the number of employees in a company and hence null hypothesis was rejected.

**3:**

****

In the multiple comparisons table, the values that are less than .05 are significantly different. The values that are significant is also indicated by the asterisks symbol. From the above table, only group 3 and group 1 are significant. That means companies having employees range between 10 to 49 and 250 above have the significant number of trademarks.

**4:**



Means plot gives the mean scores for various categories. The number of employees ranging from 10 to 49 as category 1 have the highest number of trademarks compared with 250 and above employees with lowest.

**5: EFFECT SIZE CALCULATION:**

The effect size is calculated manually by dividing sum of squares between groups with total.

Eta Sq.= = 0.10

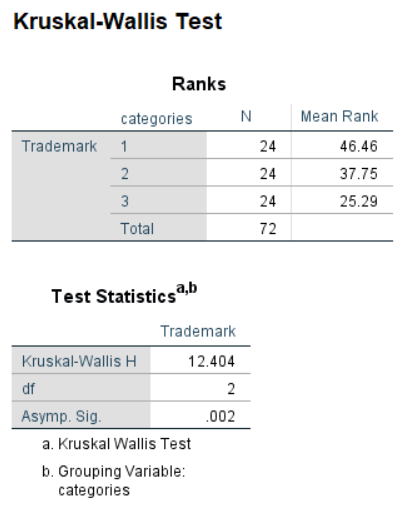
In Cohen’s terms, it is considered as medium effect (0.6 is considered as medium effect).

**CONCLUSION:**

One-way analysis of variance was conducted to find out the influence of number of employees in a company on the number of trademarks owned by company. Employees were divided into three categories (category 1 - 10 to 49 employees; category 2 – 50 to 249 employees: category 3 consists 250 and above employees). Category 1 and 3 are statistically significant. The effect size is 0.10. Category 1 ( M=1905, SD=3271) and category 3 (M=272, SD=107) are statistically different when compared with category 2.

**KRUSKAL WALLIS TEST:**

As homogeneity of variances assumption is not satisfied, Kruskal test is performed to find the variance between the different groups. Assumptions for Kruskal test are same as anova above. It is divided into different ranks with highest at the top corresponding to the scores of continuous variable. The asymptotic significance is .002 which is less than .05 indicating that difference is present across the three categories of employees.



**CONCLUSION:**

Kruskal test shows that there is significant difference among the different categories of employees and p=.002.

**(3) STATISTICS FUNDAMENTALS**

**(i) INDEPENDENT SAMPLES T-TEST:**

By performing independent sample t test, significant difference that is present in the mean scores of two groups can be found out. In our model, test is performed to compare the mean of student GPA which is a continuous variable with categorical variable gender.

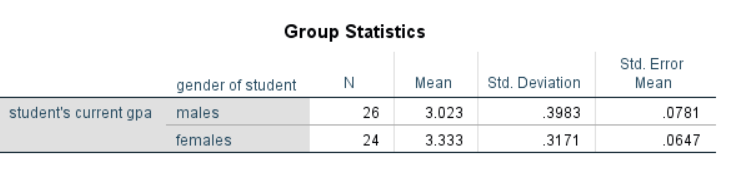
**OBJECTIVE OF THE ANALYSIS:**

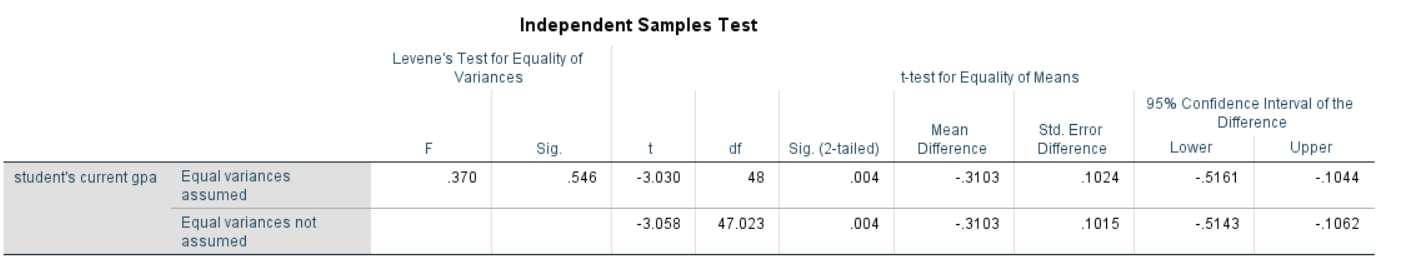
To find if there is any significant difference in the mean of current GPA scores in both the genders.

**DATASET USED:**

The dataset available on Moodle consists of fifty student data with eighteen variables. Current GPA score which is continuous is taken as dependent variable and gender which is categorical variable is taken as independent variable.

**ASSUMPTIONS:**

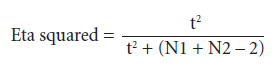
****



As the significance value for Levene’s test is greater than .05, Equal variances assumed column was taken into consideration. Therefore, data does not violate assumption. Then coming to 2-tailed significance value, it is 0.004 which is less than ideal value 0.05 hence proves there is a significant difference between in current GPA of both the gender students. Hence significance value is less than .05 between mean of current GPA, null hypothesis is rejected.

**CALCULATING EFFECT SIZE:**

It indicates the magnitude of differences between the groups by and taking events into consideration that are occurred not only by chance. One of such measure is calculating Eta squared which gives variance present in dependent variable explained by independent variable. The below formula is used for calculating the Eta squared where N1 and N2 are the number of samples in the group.



Eta sq.= (-3.03)2

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ = 0.16

(-3.03)2 + (26+24-2)

The effect size is 0.16 which is greater than large effect value (0.14). Hence 16 percentage of variance in current GPA scores was explained by gender.

**CONCLUSION:**

This test was done to compare the current GPA scores for male and female students. There was significant difference in male (M=3.02, SD=0.39) and female students (M=3.33, SD=0.31) and t(48)=-3.03.2-tailed significance value is 0.004 (<.05). The magnitude of differences in means was high (Eta. Sq=0.16).

**(ii) CHI SQUARE INDEPENDENCE TEST:**

Chi square test is performed on two categorical variables in the given population to determine the level of statistical significance associated between the categorical variables. It is a type of non-parametric test.

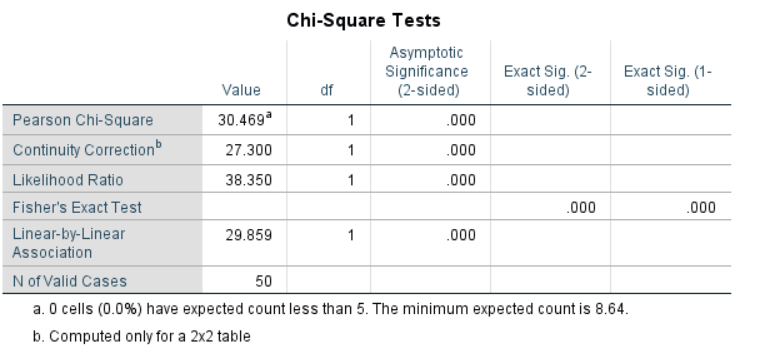
**OBJECTIVE OF THE ANALYSIS:**

To find the proportion of male watching movies is same the female proportion.

**DATASET USED:**

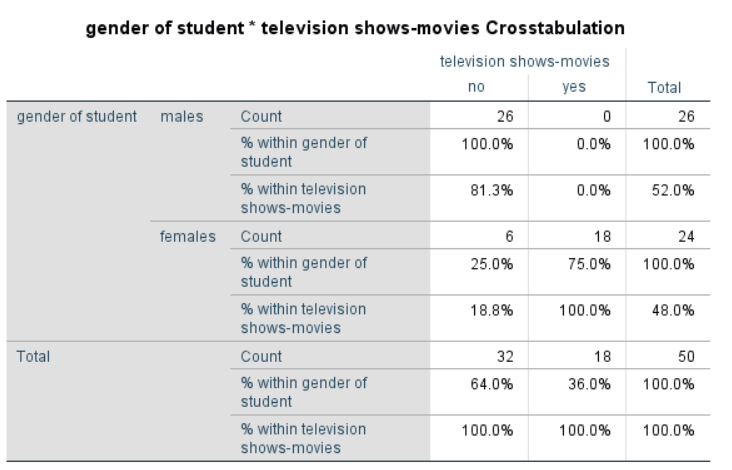
The same dataset which is used above is taken for analysing the association between males and females in watching movies. Two categorical variables are taken into consideration that includes gender (Male or Female) and tvmovies (Yes or No).

**ASSUMPTIONS:**

****

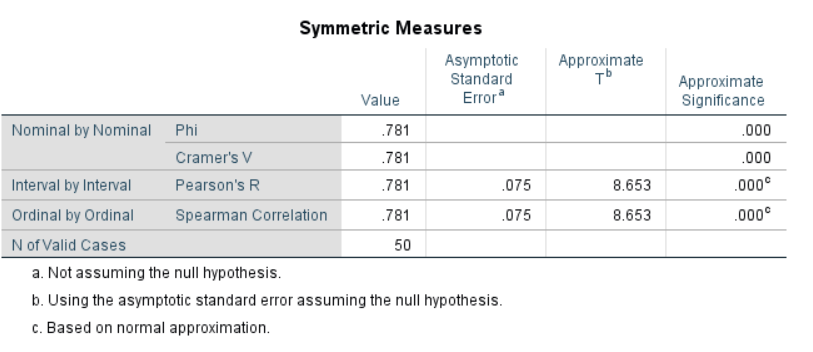
The random samples were taken. The minimum cell frequency should be five or greater. In this case, it is zero cells indicated that assumption is not violated. The Asymptotic significance values are less than .05 which shows that number of males watching movies is significantly different from females. In this model null hypothesis is rejected as number of females and males watching movies is different. The default level of significance (alpha) used is .05.

**OUTPUT ESTIMATION:**



75% of females watch movies whereas no male students watch movies. Over all 36% of population watch movies. As 2-tailed significance value is less than .05, hence null hypothesis is rejected.

**EFFECT SIZE:**

****

For 2 by 2 tables, the effect size is calculated by using Phi coefficient that ranges from 0 to 1. In this case Phi coefficient is 0.781 indicating strong association between variables.

**CONCLUSION:**

The Chi square test for independence indicates that there is a significant association between gender and movies.x2 (1,n=50)=27.30 and phi=.781.