

**STATISTICS FOR DATA ANALYTICS CONTINUOUS ASSESSMENT- PROJECT- 2 REPORT**

Kamesh Munuswamy

x19199562

MSc in Data Analytics- Batch A

National college of Ireland, Dublin

[X19199562@student.ncirl.ie](file:///C:\Users\Kamesh\Desktop\Statistics\Project\Logistic%20regression\X19199562@student.ncirl.ie)

**MULTINOMIAL LOGISTIC REGRESSION MODEL**

**Objective of the analysis:**

The idea of this analysis is to predict the current pandemic COVID-19 threat perceptions using multinomial logistic regression in IBM SPSS tool.

**Data insight:**

The data used in this analysis is browsed from Pew research centre (<https://www.pewresearch.org/>) by a survey Wave 64 conducted by Public Affairs (ipsos) from March 19th ,2020 to March 24th 2020. The dataset has a total record of 11,537 out of which 1071 cases (approximately 9%) were randomly sampled from the original cases. The dataset can be accessed using the below provided link,

*Dataset source link* **-** <https://www.people-press.org/dataset/covid-19-late-march-2020/>

**Data Pre-processing and Transformation Process:**

In this implementation, we have examined the dataset through various analysis to derive the predictors and the dependent variable of perceptions of COVID-19 virus threat. The independent variables and dependent variable are classified accordingly to implement multinomial logistic regression.

***Dependent variable encoding:*** The target variable, “Covid\_19\_threat” is encoded has 0 denotes covid- 19 is not a threat, 1 denotes covid- 19 is s minor threat and 2 denotes covid- 19 is a major threat.

***Independent variable encoding:*** The predictors in this model includes gender bias which is been named has “Male\_id” encoded has male = 0 and female = 1. The second independent variable is denoted has “Covid\_news\_impact” which carries the strong information and news related to Covid-19 pandemic.Covid\_news\_impact has four categories ranges from 0= no impact to 3= massive impact. This data includes different age categories to analysis the impact of Covid-19 across different age category. “Age\_category” is encoded has 0=18-29, 1=30-49, 2=50-64, 3=65+. Furthermore, we have recoded the age\_category variable into three dummy variables, ensemble the comparison between the different age group were 18-26 age group (serves as a baseline category which is encoded as 0 among all the dummy variables). Whereas, corresponding age groups are names has “Age30to49”, “Age50to64” and “Age65+” with encoded 1 has their dummy value.

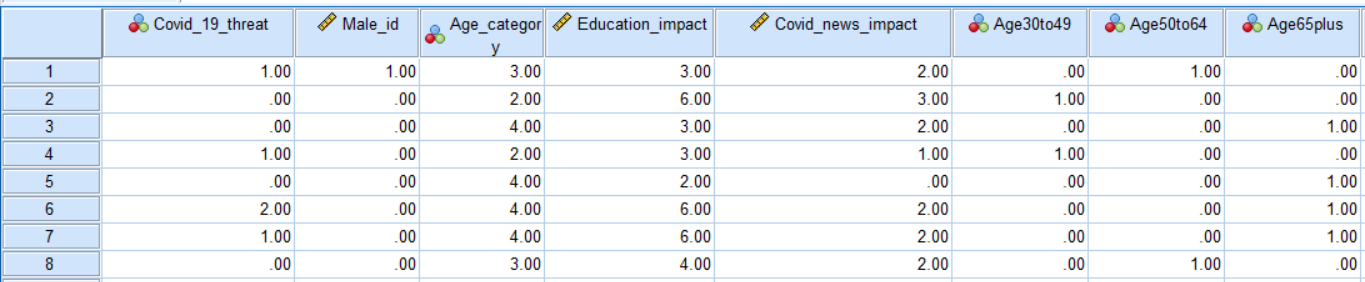
**Variables Insight:**

The pre-processed data has total of eight variables out of which, one target variable and seven predictors (3 dummy variables).

*Dependent variables* *→* Covid\_19\_threat (Nominal measure)

*Independent variables →* Male\_id, Age\_category, Education\_impact, Covid\_news\_impact, Age30to49, Age50to64, Age65plus.

**Sample data snapshot:**



**Model Assumptions:**

* First, Multinomial Logistic Regression requires the dependent variable to be a nominal measure.
* Second, the logistic regression requires that the assumptions to be independent of each other. In certain words, the observations should not be based on recurrent measurements or replicated data.
* Third, the logistic regression assumes that the multiple independent variables have little to no multicollinearity. This means that the independent variables should not be too strongly related.
* Fourth, logistic regression claims that independent variables and log odds are linear. However, this analysis does not necessarily require linear correlation between the dependent and independent variables.
* Finally, logistic regression ideally requires huge sample data in order to perform accurate prediction as possible.

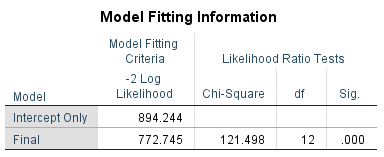
**Model Implementation in IBM SPSS Statistics:**

* In order to build a Multinomial logistic regression model, pre-processed metadata is imported into the SPSS statistical software and further evaluation on the output is interpreted.
* From the SPSS software select the mentioned option to run Multinomial Logistic Regression model. **Analyze 🡪 Regression🡪 Multinomial logistic regression**
* Model is built by choosing related assumption features in order to validate the overall model fit.
* Model assumptions are validated using the following features in SPSS,

Pseudo R-square, Classification table, Goodness-of-it, Model fitting information Likelihood ratio tests.

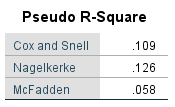
**Model Output Interpretation:** Following validation of results obtained my multinomial logistic regression is demonstrated below,

**Model fitting information:**



This segment of the output contains the outcome of the likelihood ratio of the chi-square test. On the basis of the LR-test, the model containing a complete range of predictors showcase significantly good fit compared with null model **[LR μ2(12) = 121.498, p<.001].** we can infer, at least one population slope is non-zero.

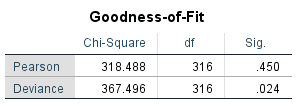
**Pseudo R-square:**



Below is the formula for the McFadden Pseudo R-square (based on Pituch & Stevens, 2016) where you construct the ratio of the variance in model deviations to the null and full model to the model deviation for the null model.

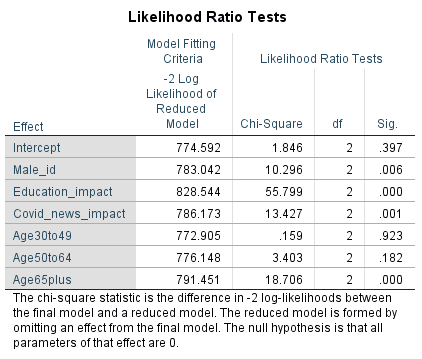
Dependent on McFadden's, we could conclude that the full model featuring our predictors reflects a **5.8%** improvement in fit compared to the null model.

**Good-ness of fit:**



This is an additional chi-square goodness of fit test for the model. When non- significant, they provide more proof of a well-fitting model. The outcomes of these experiments do not agree often with each other, as we can see here. The Pearson chi-square test implies a better fitting model (p=.450), while the Deviance chi-square shows the reverse (p=.024).

**Likelihood ration- Tests:**

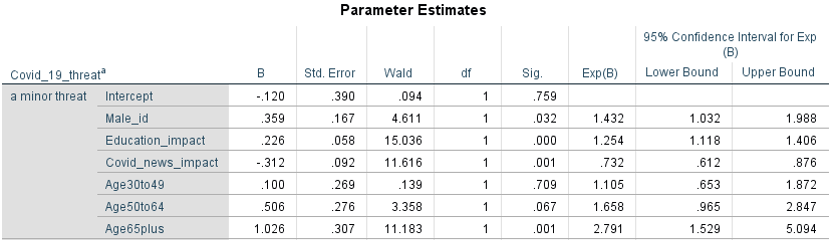


This likelihood ratio chi-square tests can be considered as omnibus tests to understand the impact of each independent variable in the model.

**Parameters Estimates:**

This section helps us to decide which of the predictors significantly predicts, whether an individual fall into the 'minor threat' category (i.e. the comparison group) versus the 'not a threat' (i.e. the baseline) category.

**A Minor threat category:**

****

*Male\_id-* The predictor is positive and significant (B=.359, Std. Error =.167, p=.032).  The slope is defined as follows: 'The log-odds of category 'minor threat' (relative to the category 'not threat') for males are estimated to be.359 points higher than females.' The odd ratio suggests that the odds involved in the category 'minor threat' for males is 1.432 times than that of odds of female. This finding shows that ***males (coded 1) are at higher probability of considering the disease as a 'minor threat' (and at lower risk of considering it to be 'not a threat') than females.***

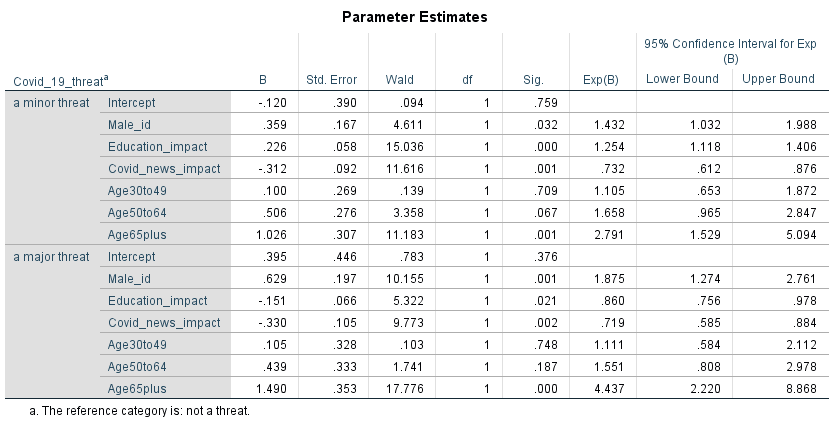
*Education\_impact-*The variable is positive and significant (B=.226, Std.Error = .058 p<.001). Once again, the likelihood ratio is greater than 1. It indicates that ***people with more education are at greater risk of assuming that Covid-19 is a 'minor threat' (relative to 'not a threat') to individual health than those with less education.***

*Covid\_news\_*impact- The predictor is negative and significant (B=-.312, Std. Error =.092, p<.001). The odds ratio is .732, meaning that with the rise in the scores on this indicator, the chances of falling into the category of 'minor threat' are changing by a factor of.732.***People who firmly believe the Covid-19 impact news are at a lower risk of believing that the disease is a 'minor threat' (and at a higher risk of believing that it is 'not a threat').***

*Age\_Category-* Since the ‘Age\_category’ variable was categorized in our initial dataset, considering the younger category (18-29) as a baseline category by which comparison among different age categories is carried out. From the result, only the ‘Age65plus’ is statistically significant (p=.001). This demonstrates that ***people aged 65+ are predicted to be at a higher risk of falling into the 'minor threat' group (and at a lower risk of falling into the 'not a threat' group) relative to people aged between 18-29.***

**A Major threat category:**

Likewise, the values interpreted for a ‘minor threat category’ above. Predictors are evaluated for ‘a major threat category’ by comparing the slope(B), Significant (Sig.) and Exp(B) values.



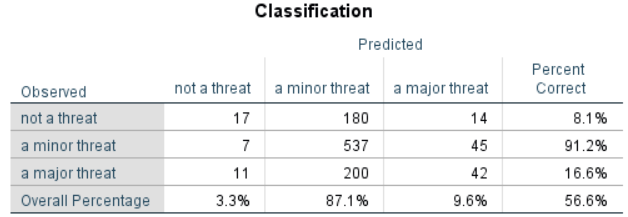
*Male\_id-* The predictor is positive and significant (B =.629, Std.Error =.197, p=.001), with an OR > 1. This suggest that ***males (coded 1) are at higher risk of considering the disease as a 'major threat' (and at lower risk of seeing the disease as 'not a threat') relative to females.***

*Education\_impact-* The predictor is negative and significant (B = -.1511, Std. Error =.0655, p=.021), with an OR < 1. This reveals that ***people with more education are at a reduced probability of believing that Covid-19 is a 'major threat' (and at a greater risk of assuming that it is 'not a threat') relative to people with less education.***

*Covid\_news\_impact-* The predictor is negative and significant (B =-.330, Std. Error  =.105, p=.002) – and OR < 1 suggest that ***people with high belief in news and knowledge about the disease are at a lower risk of believing that Covid-19 is a 'major threat' (and at a higher risk of believing that it is 'not a threat').***

*Age\_Category-* Among the dummy variables only the ‘Age65plus’ category is significant (p=.000). This indicates that ***people aged 65+ are expected to be at a higher risk of falling into the 'major threat' category (and at a lower risk of falling into the 'no threat') relative to people aged 18-29.***

Conclusion:



From the above table, the model confidently predicted that an individual will fall into the **'minor threat' category** at a rate of 537/ (7 + 537 + 45) \*100% = **91.2%**. To conclude, the model managed to attain the **overall accuracy of 56.6%** only when trying to predict using the chosen Independent variables across all three levels of the dependent variable.

**PART-B**

**One-way ANOVA**

**Definition:**

The one-way variance analysis (ANOVA) is used to evaluate if there are statistically significant differences between the means of two or more independent (unassociated) groups.

Since one-way ANOVA is an omnibus test statistic, it cannot establish which individual groups are statistically significantly different from each other; it just means that at least two groups are different from each other.

**Null Hypothesis and Alternative hypothesis:**

The null hypothesis (H0) determines the population means of two unrelated groups are equal. In other word, it denotes there is no statistically significance difference between the variables.

**H0: μ1 = μ2 = ... = μg**

The alternative hypothesis () determines the population means of two unrelated groups are not equal. It also means that there is a statistically significance difference exists between the variables.

**: not all μi (i = 1, ... *g*) are equal**

**One-way ANOVA Assumptions:**

Some expectations (or conditions) need to be satisfied to implement or execute a one-way ANOVA test. If these below mentioned criteria are not met, then the outcome of ANOVA technique may be inconsistent.

* Each sample used should an independent random sample.
* It is important to have continuous dependent variable.
* It is necessary for an independent variable consisting of two or more categorical independent group.
* The distribution of the response variables should show normal distribution.
* The data should not have much of outliers. Considering the fact that not all the outliers are bad.
* The data is required to show homogeneity of variance.

**Data and variables insights:**

For this analysis of one-way ANOVA, we have extracted the data from the given Eurostat website (<https://ec.europa.eu/eurostat>). In this dataset we have considered the vegetable production of year 2016 of 6 countries such as Sweden, Portugal, Italy, Hungary, Romania and Iceland and then merged together as a single dataset.

Variables considered,

***Independent variable*** → Countries (1= Hungary, 2 = Iceland, 3 = Italy, 4 = Portugal, 5 = Romania, 6 = Sweden)

***Dependent variable*** → Year\_2016 (continuous variable)

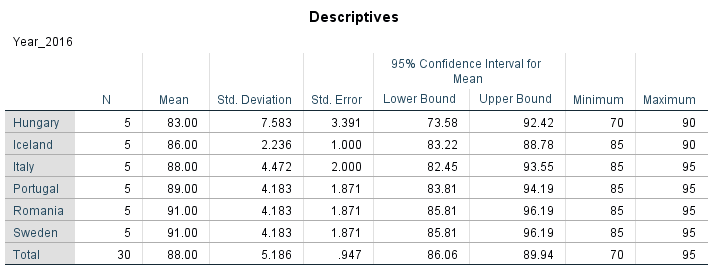
**Implementation in IBM SPSS Software:**

For this analysis we use IBM SPSS Statistic software to perform and interpret one-way ANOVA.

In SPSS software navigate through the below mentioned options to run one-way ANOVA test.

**Analyze → Compare means → One-Way ANOVA**

**Output interpretation of one-way ANOVA:**

**Descriptive table:**

The below descriptive table gives the overall statistics insights about the dependent and independent variables in terms of sample size (N), mean, standard deviation and confidence interval of mean. From this table we can infer one of the important figures of mean with respect to independent variables. It is showcase that, two countries Romania and Sweden share the highest of mean of 91 while Hungary with lowest mean of 83. This gives us the assumption that Romania and Sweden had peek vegetable production in the year 2016 meanwhile Hungary has the least production on the same year when compared with the other countries.

# **Homogeneity of variances test:**

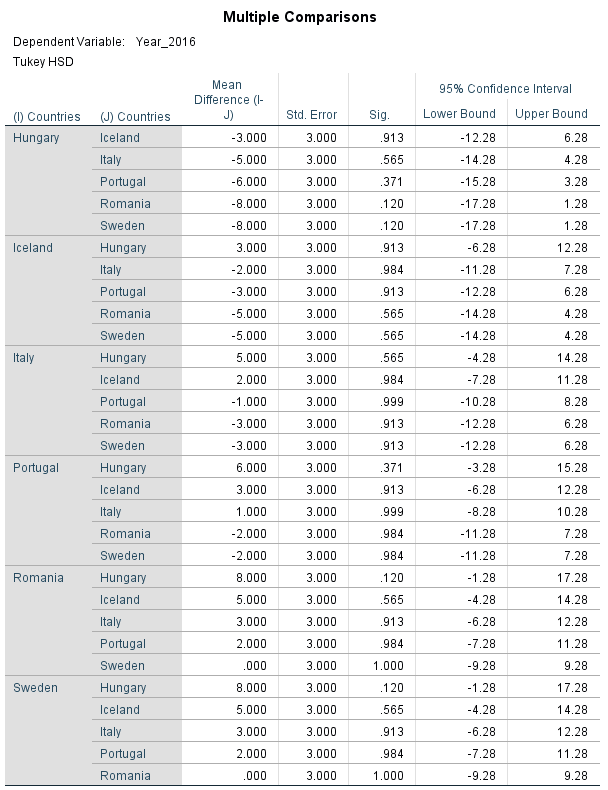
The test of homogeneity of variance comes along with one-way ANOVA in order to determine the existence of variance among the group of independent variables. From this table, ***Levene Statistic*** test is carried out to measure the variance significance. It is been identified that, **p = .488** which is greater than significance level 0.05 shows the positive result in data which resembles the sign of homogeneity of variance.

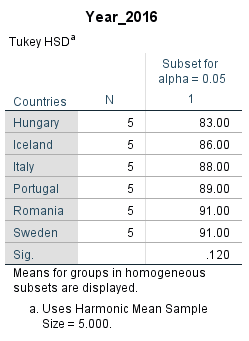
# **ANOVA table Interpretation:**

The ANOVA table with one-way ANOVA statistics defines the statistical significance of group means. It is been observed that Significant value = 0.096 which is quite larger than 0.05 significance level. So, we can infer that there is **no significance difference in mean** vegetable production among the different countries.

**The Multiple comparison Table:**

From this below mentioned table, each group is compared within each other to check the difference in mean. In this test we have performed Tukey post hoc test as a part of post hoc test, as we check how values of the countries differs from each other. As we can see that all the values have a significance level of p > 0 .05, we can say that there is no difference between the countries.



**Homogeneous Subsets Table:**

As a part of Turkey HSD test, homogeneous subsets table gives the means of groups in using the harmonic mean sample of sample size = 5.000. It is been seen that vegetable production of countries of different groups are not significantly different from each other as they all grouped in a same category 1. N denotes the mean sample size of the different group where N= 5 as we have total of 30 records with 6 different countries.

**Conclusion:**

From the above implementation of one-way ANOVA, it is observed that there is no statistically significance difference among the groups with F= 2.133 and p = 0.096. The Turkey post hoc demonstrated that vegetable production was statistically significantly higher among different country groups with respect to p value > 0.05. This test also showcased the the homogeneity of variance assumption to be positive**. Thus, it is seen that ANOVA table with p =0.096, which is p > 0.05, we fail to reject null hypothesis, meaning that all population means are equal.**

**PART C – Fundamentals of Statistics**

**INDEPENDENT SAMPLE T-TEST AND CHI-SQUARE TEST**

**INDEPENDENT SAMPLE T-TEST**

**Objective:**

The motive of this study is to perform Independent sample t-test on the provided College student data. In this analysis, Independent sample t-test is performed to compare the height of male and female students.

**Definition:**

**Independent sample t-test** is also named as ‘*Two-sample t-test’* or ‘*Students t-test’,* is an inferential statistical test which decides whether there is a statistically significant difference exist between the means of two different groups.

**Null and alternative hypotheses of independent sample t-test:**

The null hypothesis (H0) determines the population means of two unrelated groups are equal. In other word, it denotes there is no statistically significance difference between the variables.

The alternative hypothesis () determines the population means of two unrelated groups are not equal. It also means that there is a statistically significance difference exists between the variables.

**Test Assumptions:**

In order to perform the independent sample t-test, following assumptions are to be met,

* One independent, categorical variable that has two levels or groups (e.g., male/female).
* One or more continuous dependent variable.

**Data and variables insights:**

For this test, College student data which is provided is used. This dataset contains the various characteristics information about the college students of both groups (male and female) with sample size of 50 records of different ages. This test is performed to determine height comparison among the male and female students.

Variables considered,

***Independent variable*** → Gender of the student (1= male, 2 = female)

***Dependent variable*** → Height (continuous variable)

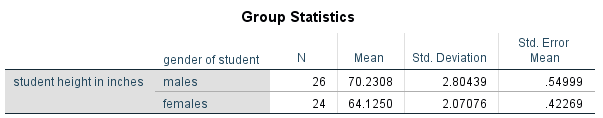
**Implementation in IBM SPSS Software:**

For this analysis we use IBM SPSS Statistic software to perform and interpret Independent sample t-test.

In SPSS software navigate through the below mentioned options to run the Independent sample t-test. **Analyze → Compare means → independent-sample t-test**

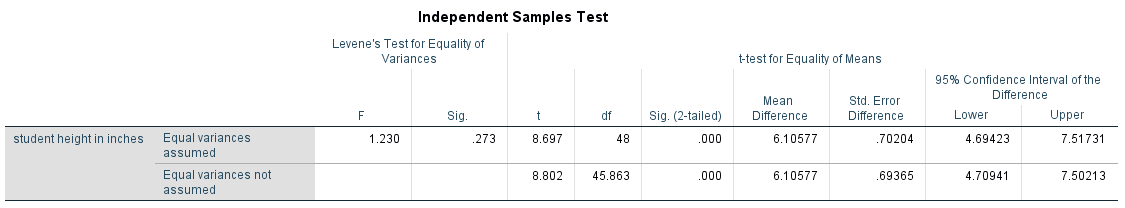
**Test output interpretation:**

**Group Statistics:**



The above table shows that, male group (N = 26) was associated with student height in inches of sample mean M= 70.230 (SD = 2.804). By comparison, female group (N = 24) was 1(SD = 2.070). Descriptively, it indicates that male students scored higher on average than female students on the basis of student height. The considered ***significant level is 0.05.***

**Independent sample t-test of Null/Alternative hypothesis:**



This component of the output is being used to determine whether or not we maintain or reject the null hypothesis (no difference in population mean). The test involves the use of theoretical sampling distribution called 't-distribution.' Basically, we equate the difference in means measured in our data with the distribution of the difference in means on the hypothesis that there is no mean population variation. The independent sample t-test associated with statistically significant effect of ***t- value t(48) = 8.697.*** From the above table, we can see that ***p value p =.000*** which is ***less than α =.05*** where alpha (α) is referred to significant level. As a result, we ***reject the null hypothesis (H0)*** and we infer **there is statistically significant difference between the male and female student.** Additionally, the assumption of homogeneity of variance was tested through ***“Levene’s Test”*** ***F(48) = 1.230, p = .273.*** The value of Levene’s test (.273) suggests no violation of homogeneity of variance assumption (i.e., homoskedasticity).

**CHI-SQUARE TEST OF INDEPENDENCE**

**Objective:**

The Objective of this analysis is to perform Chi-square test for independent using the provided college student data. In this study, Chi-square test is performed to understand the association of male and female student on watching the sports program in television.

**Definition:** The chi-square test for independence is also named as “***Pearson's chi-square test”*** or “***the chi-square test of association”***, performed to discover any relationship between two categorical (nominal, ordinal or combination of both) variables.

**Null and alternative hypotheses of Chi-square test of independence:**

In this test, the null and alternative hypothesis can be stated as follows:

***Null hypothesis*-** Corresponds to there is no relationship between two variables in the population.

***Alternative hypothesis-*** Corresponds to non-zero relationship between the two variables in the population (i.e., association between the two variables in the population).

**Test Assumptions:**

In order to perform the Chi-square test, following assumptions to be checked,

* Both the variables should be measured at an ordinal or nominal level (i.e. categorical data).
* The two variables must consist of two or more categorical, independent groups.

**Data and variables insights:**

In this test, College student data which is been provided is used. This dataset contains the various characteristics information about the college students of both groups (male and female) with sample size of 50 records of different ages.

This test is performed to check the association between the student group (both male and female) on watching sports program in television.

***Independent variable*** → Gender of the student (1= male, 2 = female)

***Dependent variable*** → Tvsports (labelled as Television shows-sports which is a nominal measure).

**Implementation in IBM SPSS Software:**

On choosing the appropriate variables by considering the assumptions, Chi-square test is performed in SPSS software.

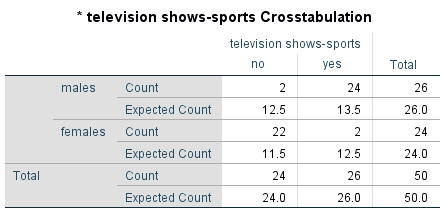
In SPSS software navigate through the below mentioned options to run the Chi-square test.

**Analyze → Descriptive statistics → Cross tabs**

The chosen variables are dragged onto the rows and columns sections. From the Statistics tab, ***Chi-square*** and ***Phi and Cramer’s V*** is selected. And the from the cells tab under count section **observed** and **expected** values is selected to infer the model prediction on the variables.

**Test output interpretation:**

**Cross tabulation:**

****

The above cross tabulation tables contain the cell frequencies and the expected count of male and female group. The observed and expected count for the student group as male and female:

Male student not watching sports show (observed = 2, expected = 12.5)

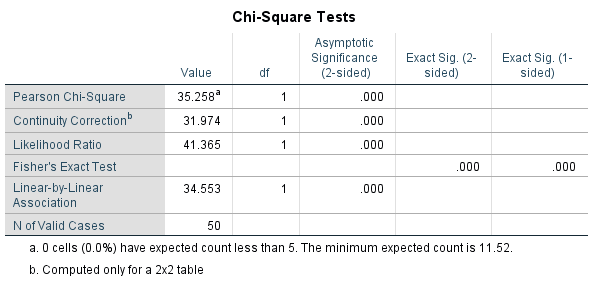
Male student watching sports show (observed = 24, expected = 13.5)

Female student not watching sports show (observed = 22, expected = 11.5)

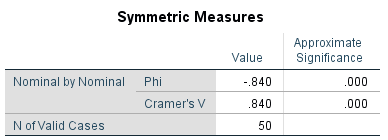
Female student watching sports show (observed = 2, expected = 12.5)

The difference between the observed and expected frequencies are unstandardized residuals.

**Pearson’s Chi-square test of Null/Alternative hypothesis:**



From the above table, the Pearson Chi-square test clearly demonstrate that **there is a statistically significant association between student group (male and female) in watching sports show in television,** ***Pearson’s χ²(1) = 35.258, p = 0.001*** which is less than significant level value = 0.05. So, we ***reject the null hypothesis (H0).*** The likelihood test is performed based on the likelihood theory. This test comes in handy when sample size is smaller. Unsurprisingly, the likelihood ratio also show the statistically significant relationship between student group (male and female) in watching sports show in television, LR χ²(1) = 41.365, p<.001. One of the expectations when conducting a chi-square test is that none of the expected cell counts should be less than 5. In such cases of assumption violation, we can rely on the results of Fisher's Exact Test. Even though, the assumption is not violated in this analysis the result shown by this test is also statistically significant association (p<.001).

**Phi and Cramer’s V – Effect size measure:**

Cramer’s V value is considered to be a measure of effect size, where value of 0 denotes no association among the two variables. According to the benchmark of chi-square tests for goodness of fit the value of **Cramer’s V= .840 indicates a larger effect.**

**References:**

Vittinghoff, E., Shiboski, S. and Glidden, D., 2014. *Regression Methods in Biostatistics*. 2nd ed. [Place of publication not identified]: New York: Springer.

Statistics.laerd.com. 2020. *One-Way ANOVA In SPSS Statistics - Step-By-Step Procedure Including Testing of Assumptions*. [online] Available at: <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php> [Accessed 1 May 2020].

Statistics How To. 2020. *Independent Samples T Test (Unpaired Samples): Definition, Running - Statistics How To*. [online] Available at: <https://www.statisticshowto.com/independent-samples-t-test/> [Accessed 1May 2020].