National College of Ireland

Statistics for Data Analytics

Statistical Analysis using Logistics regression, ANOVA and Fundamentals of Statistics models

Submitted by:

Victory Chimamaka Uwaoma
MSc. Data Analysis
National College of Ireland
Dublin, Ireland
x19210931

Submitted to: Tony Delaney

CONTENT

1. AIM AND OBJECTIVES

This research project aims at performing Logistics regression, ANOVA and Fundamentals of Statistics models which includes Independent Sample t test and Chi-square test on three distinctive datasets. The software used to perform these analyses is the SPSS version 25, the logistics data on paid Family and medical leave in the United states was taken from Pew Research Center and the ANOVA data on the percentage of education attained by population from European Union data.

2. RESEARCH QUESTIONS

a. **Logistics regression**
what is the probability that a person is working or not from given the predictors in the model (household head and marital status)?

b. **ANOVA (one-way)**
is there a difference in education attained for people across different age groups?

c. **Fundamental statistics models**
- **Independent Sample t test**
Comparison of the current CGPA by male and female students

- **Chi-Square test for independence**
Is gender (male/female) associated with the television show movies (yes or no)?

3. DATA SOURCES

The first data for Logistics regression analysis was downloaded from Pew Research Center on paid family and medical leave in the United states https://www.pewsocialtrends.org/dataset/family-and-medical-leave-study/. The data comprises of 249 columns of which 4 columns were chosen for the analysis, (PPWORK, PPHHHEAD, PPMARIT and PPAGE). It has a sample size of 7963.

The second dataset for the one-way ANOVA is gotten from European Union data https://ec.europa.eu/eurostat/data/database. The data downloaded contains 5 columns which includes (location, year, sex, age and edu_attained) on the percentage of education attained by population, with a total of 210 samples. Both datasets were in .sav and .csv respectively, an acceptable format which SPSS supports.

3.1 DATA CLEANING AND TRANSFORMATION

The cleaning and transformation process were handled manually by selecting the appropriate fields to make the data fit the logistics regression and ANOVA model. The Logistics and ANOVA datasets were cleaned and transformed below:

Logistics dataset transformation

- The PPWORK was dummy coded into two categories, working and notWorking. The PPWORK originally had 6 categories “as a paid employee”, “self-employed”, “on temporary layoff from a job”, “looking for work”, “retired” and “disabled”. “as a paid employee” and “self-employed” were re-encoded as working with a value of 1(true) while all other values as 0(false) which is notWorking.
- Also, PPMARIT was re-encoded into married and unmarried. It originally had 6 categories, “Married”, “Widowed”, “Divorced”, “Separated”, “Never married”, “Living with partner”, again, Married was encoded as 1(true) for married and all others 0(false) as unmarried.

ANOVA dataset transformation

- Before downloading the dataset, I filtered based on year (2019), gender, location and year.
- I opened the datasets using excel and did some dummy variable for age. Ages between 20 – 24 was encoded as 1, ages between 25 – 34 as 2 and ages between 35 – 44 as 3. Did the same for sex, male as 1 and female as 2.

3.2 VARIABLES

The table below shows the summary of the variables of logistics regression and one-way ANOVA analysis as well as the variables for the fundamental statistics.

Logistics Regression

Table 1: Summary of the variables used in the Logistics Regression analysis

Variable Name	Type	SPSS Measure	Unit of Measurement	Class (Variable)
PPWORK	Numeric	Nominal	Working & notWorking	Dependent (Categorical)
PPHHHEAD	Numeric	Nominal	Yes & No	Independent (Categorical)
PPMARIT	Numeric	Nominal	Married & Unmarried	Independent (Categorical)

NB: PPWORK shows the current employment status of a person. The PPHHHEAD shows if the person is a household head and the PPMARIT shows the marital status of the person

ANOVA (One-way)

Table 2: Summary of the variables used in ANOVA(One-way) analysis

Variable Name	Type	SPSS Measure	Unit of Measurement	Class (Variable)
Location	String	Nominal	Countries	—
Year	Numeric	Scale	Year	—
Sex	Numeric	Nominal	Male & Female	—
Age	Numeric	Nominal	Years	Independent (Categorical)
Edu_attained	Numeric	Scale	Percentage	Dependent (Continuous)

NB: Age shows the age group to which a person belongs. The Edu_attained shows percentage of education attained in level 5 - 8

Independent Sample t test

Table 3: Summary of the variables used in the Independent Sample t test

Variable Name	Type	SPSS Measure	Class (Variable)

currgpa	Numeric	Scale	Dependent
gender	Numeric	Nominal	Independent (Categorical)

NB: currgpa shows the cgpa of students while the gender shows if the student is either male or female.

Chi-Square test for independence

Table 4: Summary of the Chi-Square test for independence

Variable Name	Type	SPSS Measure	Class (Variable)
Tv shows (movies)	Numeric	Nominal	Categorical
gender	Numeric	Nominal	Categorical

4. METHODOLOGY

During this phase, the analysis was carried out using Logistics regression, ANOVA and Fundamental statistical model (Independent t test and Chi-square test for Independence). The results were recorded and interpreted accordingly.

Preliminary step for choosing the predictors

Before choosing the predictors for the logistics regression analysis, I had to carry out a preliminary step using Principle Component Analysis (PCA). I had 3 independent variables; Marital status, Household head and Gender (they were entered in this order), but only needed to use 2 for my analysis, so I used PCA to determine which predictors where suitable for the binary logistics regression. This was determined using a scree plot.

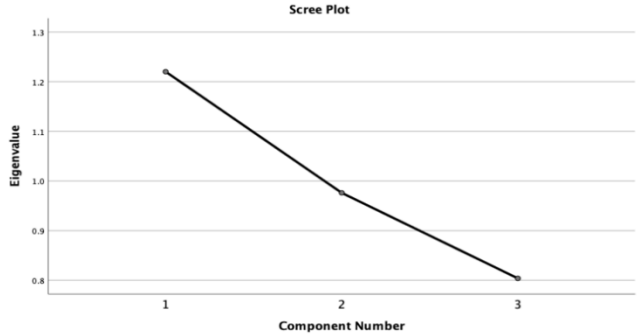

Figure 1: Scree plot

From figure 1 above, the suitable independent variables are the first 2 variables which are marital status (PPMARIT) and household head (PPHHHEAD) proved suitable.

4.1 LOGISTICS REGRESSION

Logistics regression is a statistical technique or model that is used when there is one dependent variable that is dichotomous (only 2 categories) and independent variables that can either be categorical or continuous. This research project will be discovering how accurate household head and marital status can predict the employment status of a person. A number of assumptions have to be met before Logistics regression is done to ensure that the data is appropriate for the analysis. These criteria include; sample size, outliers and linearity, multicollinearity.

a. **Sample size:** When carrying out a logistics regression analysis, it is advisable not to use a small sample size of data to avoid the issue of generalizability where the result gotten from the model cannot be used to generalize (cannot be repeated) other samples.

(Tabachnick & Fidell, 2013) provides a formula for calculating the sample size: $N > 50 + 8m$ (where N is the sample size and m is the number of independent variables). In this case, the sample data being used has a total of 7963 with 3 independent variables. Placing the following into the formula

$$7963 > 50 + 8(2)$$
$$7963 > 66$$

From the above, we can conclude that the data satisfies the assumptions of the sample size. The table below shows the case processing summary of the data.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	7963	100.0
	Missing Cases	0	.0
	Total	7963	100.0
Unselected Cases		0	.0
Total		7963	100.0

a. If weight is in effect, see classification table for the total number of cases.

Figure 2: Case Processing summary

b. **Outliers and Linearity:** The outliers and linearity of a model talks about the various parts of the score distribution and the nature that the relationship of the variables take **(Pallant, 2016)** The normal probability plot is an effective way to check the above-mentioned assumptions.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7919.754[a]	.005	.008

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Figure 3: Model summary

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.002	2	.999

Figure 4: Hosmer and Lemeshow test

The Cox & Snell R square, Nagelkerke R square and the Hosmer and Lemeshow test are used to check for the linearity of a model. The Cox & Snell R square and Nagelkerke R square have values of .005 and .008, which means 0.8% of the variances in the dependent variables can be accounted for by the predictor variables. The .999 in the Hosmer test shows that it is not statistically significant because it is greater than .5 which is a good fit model.

c. **Multicollinearity:** Multicollinearity is another assumption. It is the relation between the independent variables only. When the independent variables are highly correlated (r = .9 and above) with each other, multicollinearity is said to exist. In other to determine if the independent variables have the above-mentioned assumption, the Pearson correlation test will be done on the independent variables; marital status, household head to test for multicollinearity.

Correlations

		Marital status	Household Head
Marital status	Pearson Correlation	1	.174**
	Sig. (2-tailed)		.000
	N	7963	7963
Household Head	Pearson Correlation	.174**	1
	Sig. (2-tailed)	.000	
	N	7963	7963

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5: Pearson correlation between Marital status and Household head

Summary for Multicollinearity (using Pearson correlation)

The Pearson correlation should not be greater than .9, therefore, the figure above shows that the independent variables have a positive correlation of .174 and so this assumption was fulfilled.

Question

what is the probability that a person is working or not working from the given predictors in the model (household head and marital status)?

The idea is to test the fit of the model relative to the null model. Statistical significance for this test will be an indicator that the model is fitting the data significantly better than a null model with null predictors.

INTERPRETATION OF THE LOGISTICS REGRESSION ANALYSIS

After the assumptions were met, logistics regression was done using a single dependent variable, PPWORK and two independent variables, PPMARIT and PPHHHEAD.

Block 0: Beginning Block

Classification Table[a,b]

Observed		Predicted: Current employment status notWorking	Predicted: Current employment status working	Percentage Correct
Step 0 Current employment status	notWorking	0	1590	.0
	working	0	6373	100.0
Overall Percentage				80.0

a. Constant is included in the model.
b. The cut value is .500

Figure 6: Block 0 Classification table

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.388	.028	2452.752	1	.000	4.008

Figure 7: Variables in the equation

The figure 6 and 7 above show null model with null predictors which has an 80% accuracy.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	42.440	2	.000
	Block	42.440	2	.000
	Model	42.440	2	.000

Figure 8: Omnibus Test

Classification Table[a]

Observed		Predicted: Current employment status notWorking	Predicted: Current employment status working	Percentage Correct
Step 1 Current employment status	notWorking	0	1590	.0
	working	0	6373	100.0
Overall Percentage				80.0

a. The cut value is .500

Figure 9: Classification table

The Omnibus test of model in figure 8 shows that it is statistically significant. The classification table in figure 9 shows that the model is 80% accurate in predicting. The model predicted 0 persons as notWorking and 6373 were predicted to be working.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Marital status(1)	.099	.058	2.901	1	.089	1.104
	Household Head(1)	-.493	.074	44.204	1	.000	.611
	Constant	1.429	.038	1435.617	1	.000	4.173

a. Variable(s) entered on step 1: Marital status, Household Head.

Figure 10: Variable in the equation

From the Variable in the equation table in figure 10, it shows that household head has an unstandardized beta weight that is a negative .493 and a positive .099 for marital status. The Odd ratio Exp(B) tells us that a value of 1 shows that there is no relation between the independent and dependent variable while a value greater than 1 shows a relationship. Therefore, household head should not be used in predicting future models.

For every one unit increase in marital status, the odds of falling into the target group change by a factor of 1.104 while for every one unit increase in household head, the odds of falling into the target group change by a factor of .611

4.2 ANOVA (ONE-WAY)

ANOVA compares the mean between groups and determine if there are any statistically significant differences between the means of two or more independent groups. Assumptions have to be met before ANOVA analysis is done to ensure that the data is appropriate for the analysis. These criteria include; Normal distribution, Homogeneity of variance and linearity.

Question
is there a difference in education attained for people across different age groups?

The null hypothesis states that the population mean from the different groups are equal

$$H_0: u_1 = u_2 = u_3$$

That is the Age group 1 = Age group 2 = Age group 3 are **equal**

The alternative hypothesis is that the means are not equal. This is done using a significance level of 0.05

$$H_1: u_1 \neq u_2 \neq u_3$$

That is the means all age groups are **not equal**

a. Normal distribution

The data has to be normally distributed before ANOVA can take place and this can be checked using a histogram graph or Q-Q plots.

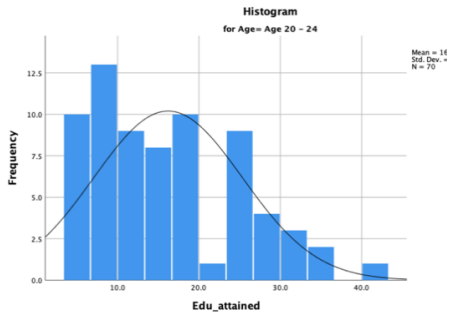

Figure 11: Histogram for Age (20 – 24)

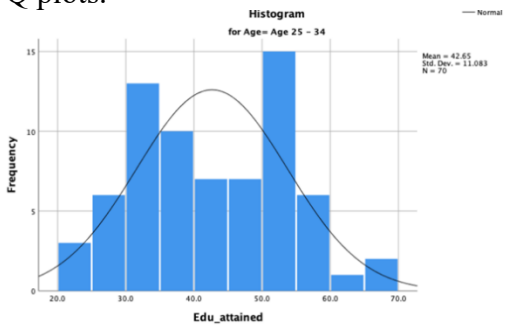

Figure 12: Histogram for Age (25 – 34) :

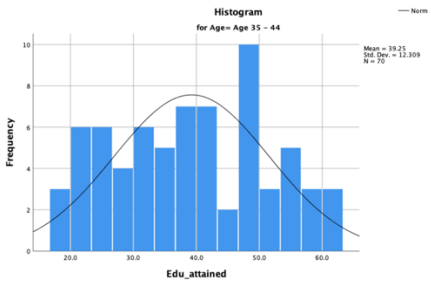

Figure 13: Histogram for Age 35 – 44)

b. Homogeneity of variance

Homogeneity of variance assumes that the variances are equal, therefore, you are hoping that the test is not significant ($p > 0.05$)

MODEL EVALUATION

The model evaluation for the ANOVA is explained below. Figure 14 shows the descriptive statistics of the Age groups. They all have equal number of samples.

Descriptives

Edu_attained

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Age 20 – 24	70	16.207	9.1278	1.0910	14.031	18.384	3.7	43.0
Age 25 – 34	70	42.647	11.0831	1.3247	40.004	45.290	21.8	69.6
Age 35 – 44	70	39.249	12.3094	1.4713	36.313	42.184	18.2	62.9
Total	210	32.701	16.0216	1.1056	30.521	34.881	3.7	69.6

Figure 14: Descriptive statistics

ANOVA

Edu_attained

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	28969.064	2	14484.532	121.489	.000
Within Groups	24679.556	207	119.225		
Total	53648.620	209			

Figure 15: ANOVA

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Edu_attained	Based on Mean	4.386	2	207	.014
	Based on Median	4.328	2	207	.014
	Based on Median and with adjusted df	4.328	2	204.784	.014
	Based on trimmed mean	4.498	2	207	.012

Figure 16: Homogeneity of variance

From the ANOVA table above in figure 15, the ($F = 121.489$, $p = .000$) is statistically significant with $p < 0.01$ and from the homogeneity of variance test, figure 16, it shows that it is also statistically significant which tells that the variance within each group are statistically different from each other but does not state which exactly. So, we can look at the Robust Tests of Equality of means below.

Robust Tests of Equality of Means

Edu_attained

	Statistic[a]	df1	df2	Sig.
Welch	144.538	2	135.794	.000
Brown-Forsythe	121.489	2	196.208	.000

a. Asymptotically F distributed.

Figure 17: Robust Test of Equality of means

From figure 17 above, the Welch statistics is statistically significant, so we reject the null hypothesis that there is actually difference in the different group means. We can use the Post-hoc test to determine which age group is statistically different from the other using the Post-hoc test table below.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Edu_attained
Tukey HSD

(I) Age	(J) Age	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Age 20 – 24	Age 25 – 34	-26.4400*	1.8457	.000	-30.797	-22.083
	Age 35 – 44	-23.0414*	1.8457	.000	-27.398	-18.684
Age 25 – 34	Age 20 – 24	26.4400*	1.8457	.000	22.083	30.797
	Age 35 – 44	3.3986	1.8457	.159	-.958	7.756
Age 35 – 44	Age 20 – 24	23.0414*	1.8457	.000	18.684	27.398
	Age 25 – 34	-3.3986	1.8457	.159	-7.756	.958

*. The mean difference is significant at the 0.05 level.

Figure 18: Post Hoc test

The [age group (20 – 24) and age group (25 – 34), p = .000], are statistically different from each other with p < 0.05 and the [age group (20 – 24) and age group (35 – 44) p = .000] are also statistically different. The [age group (25 – 34) and age group (35 – 44), p = .159]

Edu_attained

Tukey HSD[a]

Age	N	Subset for alpha = 0.05	
		1	2
Age 20 – 24	70	16.207	
Age 35 – 44	70		39.249
Age 25 – 34	70		42.647
Sig.		1.000	.159

Means for groups in homogeneous subsets are displayed.

Figure 19: Tukey HSD

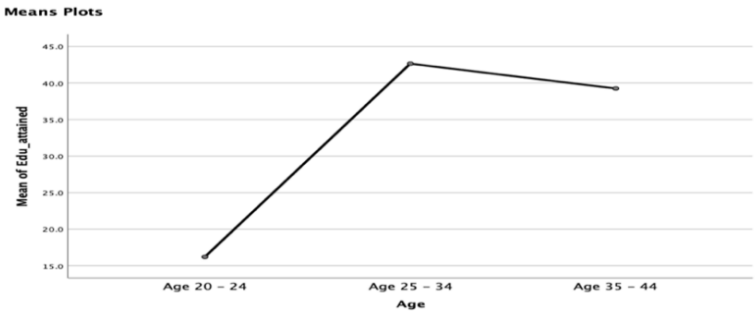

Figure 20: Means plots

The means plot above shows that age group (Age 20 – 24 yrs) has lower educational attainment than the other age groups.

INTERPRETATION OF THE ANOVA ANALYSIS

The result of the one-way ANOVA shows that there is difference in educational attainment amongst the age groups**.** Review of the means plots suggests that the younger age group (Age 20 – 24 yrs) has lower educational attainment than the other age groups.

Post-hoc comparisons using Tukey Honesty Significant Difference tests shows that the mean educational attainment for age group 1 (M =16.20, SD = 9.12) was significantly different from group 2 (M = 42.64, SD = 11.08) group 3 (M = 39.24, SD = 12.30) did not differ significantly from age group 2.

4.3 FUNDAMENTAL STATISTICAL MODELS

The fundamental statistical model includes the Independent Sample t test and the chi-square test for independence.

a. Independent Samples T test

Independent Sample t test also known as two sample t-test is an inferential statistical test which defines whether a statistically significant difference exists between the two different groups.

Question
Comparison of the current CGPA by male and female students

The null hypothesis states that the population mean from the different groups are equal

$$H_0: u_1 = u_s$$

That is the CGPA of male **is equal** to that of the female

The idea is to prove how we reject the null hypothesis and accept the alternative hypothesis.

$$H_1: u_1 \neq u_2$$

That is the CGPA of male is **not equal** to that of the female

This is done using a significance level of 0.05

To perform this analysis, we will use one categorical independent variable which is gender (male and female) and one continuous dependent variable which is CGPA of students.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
student's current gpa	Equal variances assumed	.370	.546	-3.030	48	.004	-.3103	.1024	-.5161	-.1044
	Equal variances not assumed			-3.058	47.023	.004	-.3103	.1015	-.5143	-.1062

Figure 21: Independent Sample test

INTERPRETATION FOR INDEPENDENT SAMPLE T TEST

From the above figure conducted on the Independent sample t test, the t value is -3.030 with 48 degrees of freedom and a significance level of .004. Since the significance is less than .05, we will reject the null hypothesis that the groups are equal, therefore, we can say that the CGPA of male students is different from that of the female students.

b. Chi-square test for Independence

Chi-Square test for independence is a non-parametric test. It is used to measure whether there is a relationship between categorical variables (nominal or ordinal).
There are no assumptions of homogeneity of variance, normality etc. before the Chi-Square test is be conducted.

Question
Is gender (male/female) associated with the television show movies (yes or no)?

The null hypothesis states that the population mean of gender and tv shows (movies) are not related. That means gender is independent on tv shows (movies)

The alternative hypothesis states that the population means of gender and tv shows (movies) are related. That means gender is dependent on tv shows (movies)

This is done using a significance level of 0.05

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gender of student * television shows-movies	50	100.0%	0	0.0%	50	100.0%

Figure 22: Case processing summary

gender of student * television shows-movies Crosstabulation

			television shows-movies		
			no	yes	Total
gender of student	males	Count	26	0	26
		Expected Count	16.6	9.4	26.0
	females	Count	6	18	24
		Expected Count	15.4	8.6	24.0
Total		Count	32	18	50
		Expected Count	32.0	18.0	50.0

Figure 23: Crosstabulation

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	30.469[a]	1	.000		
Continuity Correction[b]	27.300	1	.000		
Likelihood Ratio	38.350	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	29.859	1	.000		
N of Valid Cases	50				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.64.
b. Computed only for a 2x2 table

Figure 24: Chi-Square test

The Chi-Square test table in figure 24 shows that the test has not been violated because 0.0% have an expected count less than 5 and the minimum expected count is 8.64. Therefore, we can read the Pearson Chi-Square value which is 30.46 with 1 degrees of freedom. Since the p value = 0.000 (asymptotic significance) is less than .05, the test is statistically significant, therefore, we can reject our null hypothesis. That means gender is dependent on tv shows (movies).

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.781	.000
	Cramer's V	.781	.000
N of Valid Cases		50	

Figure 25: Symmetric measure

The Cramer's V ranges from 0 to 1, the symmetric measure table above shows us the effects of the association. Since the Cramer's V = .781 which is closer to 1, it tells us that there is a strong association between gender and tv shows (movies).

5. FINAL CONCLUSION

In conclusion, the binary Logistics regression which was used to answer the question of what is the probability that a person is working or not from given the predictors in the model (household head and marital status), showed that for every one unit increase in marital status, the odds of falling into the target group change by a factor of 1.104 while for every one unit increase in household head, the odds of falling into the target group change by a factor of .611. Since household head was negative, it was not a good predictor and should not be used in future model predictions.

The result from the one-way ANOVA which was used to answer the question, is there a difference in education attained for people across different age groups? showed that there is difference in educational attainment amongst the age groups**.** When the post-hoc test was used, it showed the mean educational attainment for age group 1 (M =16.20, SD = 9.12) was significantly different from group 2 (M = 42.64, SD = 11.08) and group 3. Age group 3 (M = 39.24, SD = 12.30) did not differ significantly from age group 2.

The Independent Small t test used to answer the question of comparison of the current CGPA by male and female students, Since the significance was less than .05, we rejected the null hypothesis that the groups are equal, therefore, we conclude that the CGPA of male students is different from that of the female students.

The Chi-Square test of independence used to answer the question Is gender (male/female) associated with the television show movies (yes or no)? again, since the p value = 0.000 (asymptotic significance) was less than .05, the test is statistically significant, therefore, we rejected the null hypothesis that states the population mean of gender and tv shows (movies) are not related. That's independent.

6. REFERENCES

[1] Tabachnick, B. G. & Fidell, L. S., 2013. *Using Multivariate Statistics.* 5th Edition ed. California: Pearson Education Limited.
[2] Pallant, J., 2016. *SPSS Survival Manual A step by step guide to data analysis using SPSS.* 6th Edition ed. Sydney: Allen & Unwin Academic.