National College of Ireland

STATISTICS FOR DATA ANALYTICS
CONTINUOUS ASSESSMENT 2 - REPORT

ON

PART A - LOGISTIC REGRESSION WITH PCA
PART B - ONE-WAY ANOVA
PART C – FUNDAMENTALS OF STATISTICS
(Independent Samples t test, Chi-square test for Independence)

Submitted by

Gollamudi Sarath Chandra – x19193581

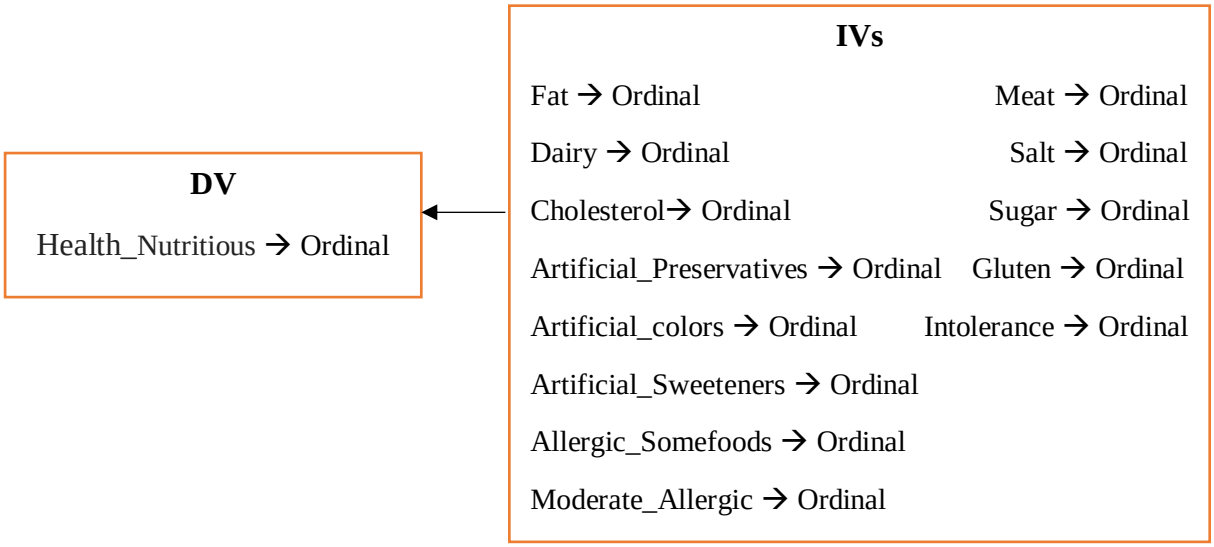
MSc Data Analytics – JAN/Group A – 2020

PART – A: Logistic Regression with PCA

Logistic regression is a statistical classification method that can be used to determine whether there are one or more independent dichotomous or continuous variables affecting the dependent dichotomous variable. Moreover, Principal Component Analysis (PCA) is a dimensionality reduction technique, which can be used to reduce many independent variables into factors or components based on the Eigenvalue.

Objective:

The main objective of this is to perform the logistic regression on the 'Public Perspectives on Food Risks dataset' to determine an outcome from IVs. On the other hand, to apply the PCA technique on the dataset to reduce the dimension of IVs and the output of this is used to do the logistic regression analysis. Below are the dependent and independent variables.



We could see in the above table there are a greater number of independent variables. In order to reduce them, at first PCA technique has applied to the dataset.

1. Dataset and Research Question:

Data Set Used: The dataset is related to the 'Public Perspectives on Food Risks dataset' which was conducted in the US. It has taken from the Pew Research Centre website. The initial dataset consists of 1000 records. After the cleaning, the dataset consists of 678 records with 13 columns with which the PCA and Logistic regression have performed.

Data Source: https://www.pewresearch.org/science/dataset/

Research Question: To validate how these IVs predict whether if the people take healthy & nutritious food.

2. Principal Component Analysis (PCA) and Assumptions Verification:

To identify the smaller number of uncorrelated variables known as principal components from a larger dataset by using a technique called PCA. In other words, dimensionality reduction. These components will be taken with the Eigenvalue which is having more than 1.

Interpretation of Output:

Step-1: To verify whether the data is suitable for component analysis we need to verify the **Kaiser-Meyer-Olkin Measure of sampling adequacy (KMO) and Bartlett's Test of Sphericity** values. In the table below we could see that the respective values are >0.6 and statistically significant (p = .000). On the other hand, we can also see that in the correlation matrix table there are more variables with the value >0.3. This means that the actor analysis is appropriate.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.808
Bartlett's Test of Sphericity	Approx. Chi-Square	1579.372
	df	78
	Sig.	.000

Correlation Matrix

		Fat	Meat	Dairy	Salt	Cholesterol	Sugar	Artificial_Preservatives	Artificial_colors	Artificial_Sweeteners	Gluten	Allergic_Somefoods
Correlation	Fat	1.000	.210	.105	.294	.410	.353	.237	.244	.194	.120	
	Meat	.210	1.000	.191	.188	.227	.221	.169	.174	.131	.162	
	Dairy	.105	.191	1.000	.030	.146	.172	.231	.191	.155	.236	
	Salt	.294	.188	.030	1.000	.291	.337	.222	.187	.241	.118	
	Cholesterol	.410	.227	.146	.291	1.000	.339	.278	.292	.236	.211	
	Sugar	.353	.221	.172	.337	.339	1.000	.359	.332	.289	.221	
	Artificial_Preservatives	.237	.169	.231	.222	.278	.359	1.000	.683	.543	.304	
	Artificial_colors	.244	.174	.191	.187	.292	.332	.683	1.000	.483	.354	
	Artificial_Sweeteners	.194	.131	.155	.241	.236	.289	.543	.483	1.000	.239	
	Gluten	.120	.162	.236	.118	.211	.221	.304	.354	.239	1.000	
	Allergic_Somefoods	-.043	.025	.059	-.044	-.010	-.046	-.048	-.024	-.079	.062	
	Moderate_Allergic	-.080	-.002	.080	-.047	-.069	-.039	.035	.053	-.014	.067	
	Intolerance	.010	.067	.285	.035	.039	.061	.113	.091	.071	.142	

Step-2: The variable values in the Communalities table tells what percentage of variance each item explained. The variables with least values indicate that it does not fit well with the other items in the component. Below is the output screenshot of the table.

Communalities

	Initial	Extraction
Fat	1.000	.529
Meat	1.000	.379
Dairy	1.000	.613
Salt	1.000	.425
Cholesterol	1.000	.501
Sugar	1.000	.468
Artificial_Preservatives	1.000	.735
Artificial_colors	1.000	.713
Artificial_Sweeteners	1.000	.587
Gluten	1.000	.428
Allergic_Somefoods	1.000	.697
Moderate_Allergic	1.000	.257
Intolerance	1.000	.733

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues Total	% of Variance	Cumulative %	Extraction Sums of Squared Loadings Total	% of Variance	Cumulative %	Rotation Sums of Squared Loadings[a] Total
1	3.410	26.230	26.230	3.410	26.230	26.230	2.693
2	1.386	10.658	36.888	1.386	10.658	36.888	2.406
3	1.191	9.165	46.053	1.191	9.165	46.053	1.585
4	1.076	8.277	54.330	1.076	8.277	54.330	1.136
5	.984	7.566	61.896				
6	.796	6.121	68.017				
7	.770	5.922	73.939				
8	.741	5.696	79.635				
9	.653	5.026	84.662				
10	.611	4.704	89.365				
11	.569	4.377	93.742				
12	.509	3.912	97.654				
13	.305	2.346	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

In the table **Communalities** above we could see that there are some items which are having less than 0.5. In order to that, we could also see in the above table **Total variance explained** the total variance by four components is only 54.330%. The minimum total variance should be greater than 60% to do the factor analysis. To increase this percentage the items with <0.5 were removed from the scale and regenerated the output. Below are the outputs.

Communalities

	Initial	Extraction
Fat	1.000	.728
Dairy	1.000	.664
Cholesterol	1.000	.684
Artificial_Preservatives	1.000	.783
Artificial_colors	1.000	.736
Artificial_Sweeteners	1.000	.634
Allergic_Somefoods	1.000	.936
Intolerance	1.000	.737

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues Total	% of Variance	Cumulative %	Extraction Sums of Squared Loadings Total	% of Variance	Cumulative %	Rotation Sums of Squared Loadings[a] Total
1	2.614	32.671	32.671	2.614	32.671	32.671	2.383
2	1.197	14.958	47.629	1.197	14.958	47.629	1.378
3	1.066	13.329	60.958	1.066	13.329	60.958	1.701
4	1.025	12.814	73.772	1.025	12.814	73.772	1.035
5	.666	8.319	82.091				
6	.585	7.309	89.399				
7	.539	6.734	96.134				
8	.309	3.866	100.000				

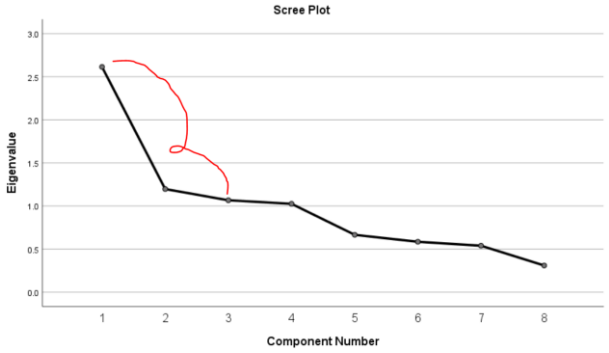
Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

From the communalities table above we can see that there are no items with <0.5 and the variance percentage also has changed. Here in the Total variance table, we have to consider only the components that are having an **eigenvalue** of 1 or more, in the table above we could

see that there are only four components recorded eigenvalues above 1 which are as follows (2.614, 1.197, 1.066, 1.025). These 4 components show a variance of 73.772 percent.

Step-3: We saw that the total four factors are extracted, so we must look at the **Screeplot** and **Component Matrix** table to confirm on the variance. Below are the output screenshots.



Component Matrix[a]

	Component			
	1	2	3	4
Artificial_Preservatives	.825			
Artificial_colors	.801			
Artificial_Sweeteners	.712		-.335	
Cholesterol	.561		.499	
Intolerance		.777		
Dairy	.402	.622	.327	
Fat	.501	-.348	.510	-.310
Allergic_Somefoods			.446	.855

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

In the graph above we could see that the first 3 components capture much more of the variance than the remaining components. On the other hand, in the **component matrix** table, only 2 items are loading on component 4 whereas the remaining 3 components are loading with more than 2 items. Based on these two observations it suggests that a three-component solution is good fit. Below are the outputs after regenerating with a fixed number of components.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.609	37.276	37.276	2.609	37.276	37.276	2.378
2	1.197	17.094	54.370	1.197	17.094	54.370	1.393
3	1.058	15.112	69.482	1.058	15.112	69.482	1.700
4	.698	9.972	79.454				
5	.586	8.373	87.827				
6	.543	7.750	95.578				
7	.310	4.422	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Component Matrix[a]

	Component		
	1	2	3
Artificial_Preservatives	.825		-.319
Artificial_colors	.802		
Artificial_Sweeteners	.710		-.349
Cholesterol	.562		.538
Intolerance		.777	
Dairy	.405	.622	
Fat	.500	-.349	.593

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

In the **Total variance** table, we can see that now the total variance explained by three components reduced to 69.482% which is more than 60% and the assumption has satisfied. However, in the **component matrix** table, each item is loading on more than 1 component which is not an ideal case. Ideally, each item should load on only one component, in order to achieve this the rotated solution method which is **Oblimin rotation** was used on three components. After applying the rotation solution below three are the new tables which we need to consider.

Component Correlation Matrix

Component	1	2	3
1	1.000	.199	.345
2	.199	1.000	.105
3	.345	.105	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Pattern Matrix[a]

	Component		
	1	2	3
Artificial_Preservatives	.873		
Artificial_colors	.842		
Artificial_Sweeteners	.813		
Intolerance		.845	
Dairy		.752	
Fat			.866
Cholesterol			.807

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Structure Matrix

	Component		
	1	2	3
Artificial_Preservatives	.883		.309
Artificial_colors	.856		.330
Artificial_Sweeteners	.792		
Intolerance		.826	
Dairy		.774	
Fat			.849
Cholesterol	.327		.825

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

From the component correlation table, we can tell the relationship between the three factors. Here, in this we could see that component three is having the value above .3, this means that we must report the Oblimin method. This will provide two output tables called **Pattern Matrix and Structure Matrix** which are shown above. In the pattern matrix table to label each

component we must look at the loading items on three components. In this the main loading on component 1 is Artificial_Preservatives item, component 2 is Intolerance and on component 3 is Fat. On the other hand, the Structure Matrix table is unique to the Oblimin output which provides the correlation between the variables and factors in this, all three are having the values greater than .3, this means that all PCA assumptions are satisfied. Finally, the output three factors are used for the logistic regression analysis.

3. Logistic Regression and Assumptions Verification:

The key aspects of the logistic regression output are highlighted in the below discussion.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	678	100.0
	Missing Cases	0	.0
	Total	678	100.0
Unselected Cases		0	.0
Total		678	100.0

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

Dependent Variable Encoding

Original Value	Internal Value
Less than half of the time	0
All of the time	1

At first, verifying the size and nature of the sample data, in the table **Case Processing Summary** above we can see that there are no missing values in our data and the sample size is as expected 678. In the next table, Dependent variable Encoding we could see that the spss coded into 0 and 1. The next output is **Block 0** which gives the results of our analysis without any independent variables used in the model. This will be used later for comparison with the predicted variables included. Below is the output classification table.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Health_Nutritious: Less than half of the time	Predicted: Health_Nutritious: All of the time	Percentage Correct
Step 0	Health_Nutritious	Less than half of the time	368	0	100.0
		All of the time	310	0	.0
	Overall Percentage				54.3

a. Constant is included in the model.

b. The cut value is .500

The overall percentage of correctly classified cases is 54.3 percent. This means that IBM SPSS guessed that all cases would have taken healthy & nutritious food less than half of the time. This is only because there is a high percentage of this answer in the table. In the next output Block, we will see the actual accuracy of the model once our predictors are included.

Another output of our analysis is **Block 1** where the actual set of predictors was tested. The below table **Omnibus Tests of Model Coefficients** gives the indication of overall how well our model performs with none of the predictors entered the model referred to as the 'goodness of fit' test.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	207.733	3	.000
	Block	207.733	3	.000
	Model	207.733	3	.000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.062	6	.061

We could see that the value is .000 which is significant ($p<0.05$) and the chi-square value is 207.733 with 3 degrees of freedom. This means that model is better than the IBP SPSS original guess. On the other hand, the results from the table **Hosmer and Lemeshow Test** also supports

our model. This is interpreted differently from the last table Omnibus. For the 'goodness fit of test' if the significance value is less than .05 then it will consider as a poor fit. In this analysis, we could see in the above screenshot the chi-square value of 12.062 with a sig value of .061 (p>.05).

The below table **Model Summary** tells us the information of the model usefulness.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	727.207[a]	.264	.353

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

Classification Table[a]

Observed			Predicted Health_Nutritious: Less than half of the time	Predicted Health_Nutritious: All of the time	Percentage Correct
Step 1	Health_Nutritious	Less than half of the time	305	63	82.9
		All of the time	112	198	63.9
	Overall Percentage				74.2

a. The cut value is .500

The **Cox & Snell R Square** and the **Nagelkerke R Square** values provide an explanation of the amount of variation in the DV explained by the model. In this analysis the values are .264 and .353, this means that between 26.4% and 35.3% of the variability is explained by this set of variables. The next output is the **Classification table** which tells us the prediction accuracy of the model. In the above table, we could also see that the accuracy is 74.2. If we compare this with the accuracy of Block 0 output 54.3 percent, we can tell that there is a good improvement in the model when the predictor variables are included. In this, we were able to correctly classify the *sensitivity* that is 63.9 percent of people choose healthy & nutritious food all the time. On the other side, the *specificity* is 82.9 percent which tells that people choose healthy & nutritious food less than half of the time.

The next output table of our analysis is the **Variables in the Equation** which tells about each predictor variable.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	REGR factor score 1 for analysis 2	.986	.102	93.073	1	.000	2.681	2.195	3.276
	REGR factor score 2 for analysis 2	.261	.091	8.274	1	.004	1.299	1.087	1.552
	REGR factor score 3 for analysis 2	.489	.097	25.270	1	.000	1.631	1.348	1.974
	Constant	-.139	.092	2.319	1	.128	.870		

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 2, REGR factor score 2 for analysis 2, REGR factor score 3 for analysis 2.

Casewise List[b]

Case	Selected Status[a]	Observed Health_Nutritious	Predicted	Predicted Group	Temporary Variable: Resid	ZResid	SResid
48	S	L**	.946	A	-.946	-4.185	-2.421
140	S	L**	.946	A	-.946	-4.185	-2.421
252	S	L**	.866	A	-.866	-2.539	-2.016
256	S	L**	.866	A	-.866	-2.539	-2.016
406	S	L**	.896	A	-.896	-2.936	-2.135
428	S	L**	.866	A	-.866	-2.539	-2.016

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

The test used for this is called as **Wald Test**. In this analysis, we have three significant values with less than 0.05. This means that the major factor influencing the DV are Factor 1 (Artificial_Preservatives, Artificial_colors & Artificial_Sweeteners), Factor 2 (Intolerance & Dairy), Factor 3(Fat & Cholesterol). The **B** values in the second column indicate the direction of the relationship. We should check whether those are positive or negative. In this case, we could see all the values are positive this tells that people accepting that they are taking healthy & nutritious food are more likely to answer 'All of the Time' to the question whether they are eating healthy and nutritious food or not. The other important values in this table are the **Exp(B)** column which gives the **odds ratio** of each independent variable. In this analysis the odds of a person answering yes, they eat artificial foods is 2.681 times greater than someone who reports All of the time than the persons saying less than half of the time. The remaining two variables are also significant with the odds ratio of 1.299 and 1.631. For each of these ratios there is a 95 percent confidence interval shown in the above table. In this case, the confidence interval of the first variable ranges from 2.195 to 3.276. So, the calculated odds ratio is 2.681, we can be 95 percent confidence that the actual value of the odds ratio lies somewhere between the given

range. Since we do not have the confidence interval with value 1; therefore, the result is statistically significant at $p<0.05$. In the table above **Casewise List,** we could see we have some samples for which the model does not fit well. Cases with ZResid values above or below 2.5 should be examined closely. So, we can consider all these are clear outliers.

Conclusion:

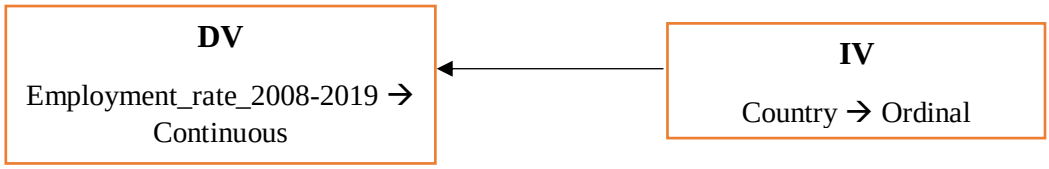
Logistic regression was performed with the output of PCA to ascertain the effects of IVs on the likelihood that participants were going to take healthy & nutritious food all the time or less than half of the time. The logistic regression model was statistically significant, x^2 (3, N=678) = 207.733, $p<0.001$, stating that the model was recognizing between the respondents. The model explained 35.3% (Nagelkerke R^2) of the variance in healthy and nutritious food and correctly classified 74.2% of cases. All 3 factors were associated with an increased likelihood of taking healthy and nutritious food.

PART – B: One-Way ANOVA

To determine whether there is any statistically significant variance between the means of three or more independent groups (variables), the One-Way can be useful. It consists of one independent variable with three or more levels and one continuous dependent variable.

Objective:

The main objective of this analysis is to find out is there any significant variance in the means of the three countries with respect to the employment.



1. Dataset and Research Question:

Data Set Used: The dataset is related to the 'Employment rates of recent graduates' which was conducted in three countries (Ireland, Greece, and Italy) which is a yearly survey. It has taken from the Eurostat website. the dataset consists of 37 records with one column.

Data source: https://ec.europa.eu/eurostat/databrowser/view/tps00053/default/table?lang=en

Research Question: To determine is there any difference in the employment rate of recent graduates between three countries from 2008 to 2019.

2. One-Way between groups ANOVA with post-hoc tests and Assumptions Verification:

Interpretation of Output:

At first, The **Descriptives** table gives information on Mean, Std. Deviation and 95% confidence interval values. It is also important to check the N value in this table to make sure all 3 levels are having the same number. Below is the screenshot of the output.

Descriptives

Employment_rate_2008-2019

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Ireland	12	78.5417	5.48211	1.58255	75.0585	82.0248	70.80	86.60
Greece	12	52.5750	8.97523	2.59093	46.8724	58.2776	40.00	68.30
Italy	12	55.0583	5.69042	1.64268	51.4428	58.6739	45.00	65.20
Total	36	62.0583	13.62821	2.27137	57.4472	66.6695	40.00	86.60

In the table above we could see in our data, all three levels are having equal records. Therefore, we are good to conduct analysis with this data.

Moving further the second table which we need to verify is the **Test of Homogeneity of Variances**, this test whether the variance in scores is the same for each of the 3 groups. In order to verify this, we need to make sure that the significance value is greater than 0.05. We could see in the table below the sign values are (.109, .155, .159, .117) all are >0.05, this means that we have not violated the homogeneity of variance assumption.

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Employment_rate_2008-2019	Based on Mean	2.370	2	33	.109
	Based on Median	1.976	2	33	.155
	Based on Median and with adjusted df	1.976	2	26.393	.159
	Based on trimmed mean	2.291	2	33	.117

Thirdly **ANOVA** table tells the sum of squares of both between groups and within groups with the degrees of freedom. The main point we need to see in this table is the Sig value. If this value is below 0.05 this tells us that there is a significance difference somewhere among the mean scores on the DV for the 3 groups. Below is the output of ANOVA.

ANOVA

Employment_rate_2008-2019

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4927.607	2	2463.803	51.692	.000
Within Groups	1572.881	33	47.663		
Total	6500.487	35			

In the table above we can clearly see that the Sig value is less than 0.05 this shows there is some difference. To verify the statistical significance of the differences between each pair of groups we need to look at the **Multiple Comparisons** table which is the table of post-hoc tests results.

Post Hoc Tests

Multiple Comparisons

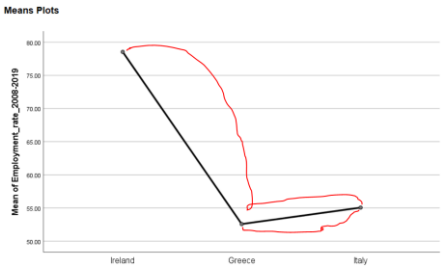
Dependent Variable: Employment_rate_2008-2019

Tukey HSD

(I) Country	(J) Country	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Ireland	Greece	25.96667*	2.81848	.000	19.0507	32.8826
	Italy	23.48333*	2.81848	.000	16.5674	30.3993
Greece	Ireland	-25.96667*	2.81848	.000	-32.8826	-19.0507
	Italy	-2.48333	2.81848	.656	-9.3993	4.4326
Italy	Ireland	-23.48333*	2.81848	.000	-30.3993	-16.5674
	Greece	2.48333	2.81848	.656	-4.4326	9.3993

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

In the table above we could see that there are some values with asterisks (*), this means that the two groups are different from each another with p<0.05. In the results presented above except (Greece-Italy, Italy-Greece) pairs, remaining pairs are significantly different from one another. That is (Ireland → Greece, Italy) group 1 differ significantly in terms of the employment rate of recent graduates. To compare the mean scores of these 3 groups we will verify the **Means Plots**, below is the output screenshot.



In the above plot, we can clearly see that **Greece and Italy** recorded the lowest employment rate, with the **Ireland** recording the highest employment rate from the year 2008 to 2019. Finally, although SPSS does not generate the **effect size** for this analysis. The eta squared value is calculated using the below formula.

Eta squared = Sum of squares between groups / Total sum of squares

Size	Eta squared (% of variance explained)	Cohen's d (standard deviation units)
Small	.01 or 1%	.2
Medium	.06 or 6%	.5
Large	.138 or 13.8%	.8

The table above is Cohen's classification of effect sizes. In this analysis, by taking the values from the ANOVA table the resulting value of **eta squared** is 0.758, which is a large effect size with respective to Cohen's d as shown in the above table.

Conclusion:

A one-way ANOVA between groups was performed to find the difference in the employment rate of recent graduates. The total participants were divided into 3 groups based on their countries (Ireland, Greece, and Italy). There was a statistical significance difference at the $P<0.05$ in the employment rate for three countries $F(2, 33) = 51.692$, $p<0.01$. On the other hand, the difference in the mean scores of between groups was large with the effect size of 0.758. The post-hoc comparisons shows the mean score of first group ($M = 78.5417$, $SD = 5.48211$) was significantly different from remaining two groups which are as follows Group 2 ($M = 52.5750$, $SD = 8.97523$) and Group 3 ($M = 55.0583$, $SD = 5.69042$).

PART – C: Fundamentals of Statistics

(Independent Samples t test, Chi-square test for Independence)

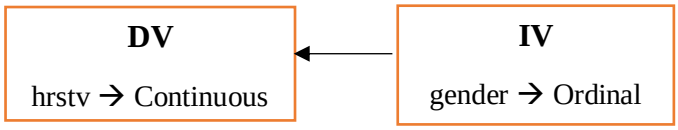
1. Independent Samples T Test:

To find out whether the population means of the groups are statistically different, the Independent Samples t Test is used by comparing the means of two independent groups.

Objective:

The objective of this analysis is to find out the variance in the amount of time spent to watch TV by comparing both groups males and females.

Dataset: The dataset used for this analysis is the CollegeStudentData survey.



Levene's Test for Equality of Variances:

This test requires the assumption of homogeneity of variance. In other words, two groups have equal variance. SPSS conducts this as a Levene's test. Below is the null and alternative hypothesis of Levene's.

H_0: $\sigma_1^2 - \sigma_2^2 = 0$ ("the population variances of group 1 and 2 are equal")
H_1: $\sigma_1^2 - \sigma_2^2 \neq 0$ ("the population variances of group 1 and 2 are not equal")

If we reject the null hypothesis, it says that the homogeneity of variance test is violated.

Interpretation of Output:

The significance level used for this analysis is α = 0.05. This means that if the result is below this α value, we will reject the null hypothesis.

Group Statistics

	gender of student	N	Mean	Std. Deviation	Std. Error Mean
amount of tv watched per week	males	26	13.04	6.328	1.241
	females	24	10.83	5.746	1.173

Independent Samples Test

		Levene's Test for Equality of Variances: F	Sig.	t-test for Equality of Means: t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference: Lower	Upper
watched per	Equal variances assumed	.821	.369	1.286	48	.204	2.205	1.714	-1.242	5.652
	Equal variances not assumed			1.291	47.990	.203	2.205	1.707	-1.228	5.638

In the table above **Group Statistics** shows the comparisons, including sample size, Mean, and Std Deviation values. We could see that there are 26 males and 24 females following with the mean values of 13.04 and 10.83. The important table in this analysis is an **independent samples test** table which is shown above. There are 3 sections which are Levene's test, t-test for equality means, and 95% confidence interval values. In the table, we could see the significance value is .369 which is above the α value. This means that we accept the null of Levene's test and tells that the variance in males and females are the same. Since the p-value is greater than α value, we need to take the **t** value of **Equal variance assumed** output for the t test and corresponding confidence intervals. This t-test provides the actual result of independent sample t-test. We could also see that the mean of females was subtracted from the mean of females. The positive t value (1.286) indicates that the mean value for the first group males is significantly greater than the females. The associated p-value is .204 which is less than the α value. In order to that, the 95% CI values are -1.242 to 5.652 which does not contain zero. This means that it agrees with the p-value of the significance test. On the other hand, the effect size for this test is (eta squared = 0.03) calculated using a formula which is small as per the Cohen's d.

Conclusion:

Since the significance value is greater than the α value (0.05) we accept the null hypothesis and there was no significant difference between the males and females in the amount of time spent to watch TV with the t value of (t_{48} = 1.286, p>0.05) with the (mean differences = 2.205, CI of -1.242 to 5.652) and the effect size was small with an eta squared value of 0.03.

2. Chi-square test for Independence:

To explore the relationship between the two categorical variables the Chi-square test for independence is useful. It is also called as Chi-Square Test of Association and it is a nonparametric test. This test is based on a contingency table (crosstabulation or two-way table),

in which each of the variables consists of more than two levels. This test can make comparisons only between categorical variables but not with the continuous variables.

Objective:

The objective of this test is to determine is there any association between the variables gender and television-shows movies.

Dataset: The dataset used for this analysis is the CollegeStudentData survey with a sample size of 50 records.

Categorical Variables

gender & tvmovies

Hypothesis:

The null and alternative hypothesis (H_0 and H_1) of the Chi-square Test of Independence are as below, and the test statistic of the chi-square is denoted by x^2.

$$H_0 = [Variable\ 1]\ is\ not\ associated\ with\ [Variable\ 2]$$

$$H_1 = [Variable\ 1]\ is\ associated\ with\ [Variable\ 2]$$

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

Interpretation of Output:

Assumption 1: The first thing we need to verify is the 'minimum expected cell frequency' which should be 5 or greater. Below is the output screenshot.

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

We could see in the above table **Chi-Square Tests** which is highlighted with red mark it shows that 0 cells have expected count less than 5. So, we have not violated the assumption.

Assumption 2: Moving further in the same table above, the important value we are interested to verify is Pearson Chi-Square value, but in our case, since it is a 2 by 2 table so we need to consider the **Continuity Correction** value which is below the Pearson value highlighted with the red mark. We can clearly see that the correction value in this analysis is 27.3 with an associated significance value of .000 which is less than the α value 0.05. So, we can conclude that our result is significant, this means that the proportion of males who claim TV does not show movies is significantly different from the proportion of females who claims the same. It shows that there is an association between the tvmovies and gender variables.

Assumption 3: To find what percentage of each gender claims the TV does not show movies we need to check the Crosstabulation table as shown below.

gender of student * television shows-movies Crosstabulation

			television shows-movies		
			no	yes	Total
gender of student	males	Count	26	0	26
		% within gender of student	100.0%	0.0%	100.0%
		% within television shows-movies	81.3%	0.0%	52.0%
		% of Total	52.0%	0.0%	52.0%
	females	Count	6	18	24
		% within gender of student	25.0%	75.0%	100.0%
		% within television shows-movies	18.8%	100.0%	48.0%
		% of Total	12.0%	36.0%	48.0%
Total		Count	32	18	50
		% within gender of student	64.0%	36.0%	100.0%
		% within television shows-movies	100.0%	100.0%	100.0%
		% of Total	64.0%	36.0%	100.0%

In this analysis, we can see that 100.0% of males claim that TV does not show movies. For females, 25% claims that TV does not show movies where the rest 75% claims that TV shows movies. To make sure that we are referring the right column we could see that total of both genders **%within gender of students** is up to 100%. To find the percentage of sample claims we must verify the values **of Total**. From these results, 64% of the sample claims TV does not show movies, 36% of the sample claims TV shows movies.

Assumption 4:

To verify the effect size for 2 by 2 table we will investigate the **phi coefficient** value in the below **Symmetric Measures table.**

Symmetric Measures

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

In this analysis since the table is 2 by 2, we need to consider the **phi** value, which is .781, this is a correlation coefficient value and can range from 0 to 1. Based the Cohen's criteria on effect sizes, we could say that in this analysis there is a stronger association between the two variables.

Conclusion:

A Chi-Square test for independence has performed on the CollegeStudentData survey dataset, the final Yates Continuity Correction value indicates that there is a significant association between the two categorical variables gender and tvmovies with x^2 (1, n=50) = 27.3, p = .000, phi = .781.