**DSCI5180 Section 006-Introduction to the Business Decision Process**

**Final Project**

**Name: Buddha Satyasriharsha EUID: 11627317**

**About the Project:**

This is a project on American Sitcom named ‘The Office’ and the relevant dataset was taken from Kaggle.com. The modified dataset consists of 5 columns and 188 rows scrapped from IMDb.The dataset contains data from all the 9 seasons.

The full dataset can be found here. **

**Applying lessons learnt from Module 1**

The ratings of each episode from the dataset are assumed to have a normal distribution with the mean of 8.237 and the standard deviation of 0.589. If an episode is randomly selected, find the probability of finding an episode with rating more than 8.5. Round the answer to three decimal places.

Solution- From the above, we know that Mean (μ)= 8.237

Standard Deviation (σ)=0.589 and X =8.5

By Substituting above values P(X>8.5) = P(Z>0.447) = 0.330.

The probability of finding an episode with rating more than 8.5 is 0.330 or 33%.

Please refer to appendix for step-by-step process for [Module 1](#Module1Reference).

**Applying lessons learnt from Module 2**

A Producer would like to determine the average number of votes for the series, He took a random sample of 31 episodes and found that the sample has a Mean of 3199.41 and a Standard deviation of 554.57. Construct a 95% Confidence Interval for average number of votes.

Solution- Since the population standard deviation is not known, we will use the t-distribution to construct the confidence interval.

For the Sample size(n)=31 we know that x̄ = 3199.41, Sample standard deviation(s)=554.57.

The degrees of freedom associated with the problem is df=n-1=31-1=30.

The t-value corresponding to 30 degrees of freedom and 95% confidence (α=0.05) is t0.025,30=2.042.

Confidence Interval = 

Substituting above values in the formula we get the 95% Confidence Interval for average number of votes to be [2995.99,3402.8]

Please refer to appendix for step-by-step process for [Module 2](#Module2Reference).

**Applying lessons learnt from Module 3**

The mean of the Viewership for entire series is 7.24 million. A critic believes that this mean is undermined and mean of the viewership is a more than given. He conducts a survey with sample size of 39 episodes and finds out the sample mean of viewership is 8.40 million and sample standard deviation is 0.97 million. If the population standard deviation is 2.066. Is there sufficient evidence to support the critic’s claim at the 0.05 level of significance

Solution-From the above, we know that x̄=8.40Million, μ= 7.24 million,

Standard Deviation (σ)=2.066 and sample size (n) =39 and Significance level α= 0.05.

The null hypothesis, H0 is the statement the mean of the viewership is 7.24 million i.e H0= μ =7.24.

The alternative hypothesis, Ha is the statement the mean of the viewership is greater than 7.24 million. i.e., Ha= μ >7.24.

This is one sided right tail test.The critical value for given value of α= 0.05 is 1.645.

The Test statistic is Z = (x̄ – μ) / (σ / √n), Substituting the above values we get the Test statistic to be 3.5063.

Since the Test statistic (3.506) is greater than the critical value (1.645), we reject the null hypothesis.

There is sufficient evidence to reject the null hypothesis and support the critic’s claim at the 0.05 level of significance.

Please refer to appendix for step-by-step process for [Module 3.](#Module3Reference)

**Applying lessons learnt from Module *5***

The following data was collected to explore how the viewership per episode and Number of votes affect the Rating of the episode. The dependent variable is rating of the episode, the first independent variable is the votes per episode and, the second independent variable is viewership (Per million) per episode. Determine if a statistically significant linear relationship exists between the independent and dependent variables at the 0.01 level of significance. If the relationship is statistically significant, identify the multiple regression equation that best fits the data. Otherwise, indicate that there is not enough evidence to show that the relationship is statistically significant.

Solution: Refer to appendix for the regression output on excel for [Module 5](#Module5Reference).

From the regression analysis, we get that the P value is 7.2943E-26,

It is less than our Alpha value of 0.01 so we can say that there is sufficient evidence at the 0.01 level of significance to conclude that the linear relationship between the independent variables and the dependent variable is statistically significant.

The regression equation is

ŷ= Coefficient of Votes\*(Votes Per Episode) + Coefficient of Viewership\*(Viewership per episode) + Coefficient of Intercept.

ŷ= 0.00030011\*(Votes Per Episode) + 0.06663101\*(Viewership per episode) + 6.90263249.

Please refer to appendix for step-by-step process for [Assumptions.](#Assumptions)

**Appendix:**

**Info of the Project:**

The dataset contains the following fields

* Season - This field helps us to know the season number
* Episode Title - This field helps us to know the Name of the Episode
* Ratings- This field helps us to know the Viewer Ratings on scale of 1-10
* Votes- This field helps us to know the Total number of votes given in survey by viewers
* Viewership- This field helps us to know the Total Number of viewers (In Millions)

The Dataset snippet is as follows*Table

Description automatically generated*

**Assumptions:**

For the entire project -Since all the data and samples has sample size more than 30, we can assume that all the data is normally distributed.

For Module 5- We Assume that there is no correlation between both the independent variables taken.

**Module 1:**

**Step-by-Step Process**

Mean (μ)= 8.237

Standard Deviation (σ)=0.589 and X =8.5

By Substituting above values P(X>8.5) =P(Z> (8.5-8.237)/0.589) = P(Z>(0.263/0.589)=

P(Z>0.447) =P(Z<-0.447) = 0.330

**Module2:**

**Step-by-Step Process**

Sample size(n)=31 we know that x̄ = 3199.41, Sample standard deviation(s)=554.57.

The degrees of freedom associated with the problem is df= 30.

The t-value corresponding to 30 degrees of freedom and 95% confidence (α=0.05) is t0.025,30=2.042

Confidence Interval = 

=[3199.41+2.042\*(554.57/SQRT(31)), 3199.41-2.042\*(554.57/SQRT(31))]

=[3199.41+2.042\*99.6037, 3199.41-2.042\*99.6037]

=[3199.41+203.3907, 3199.41-203.3907]

=[3402.80, 2995.99]

**Module 3:**

**Step-by-Step Process**

x̄ =8.40Million, μ= 7.24 million,

σ=2.066 and sample size (n) =39 and Significance level α= 0.05.

H0= μ =7.24

Ha= μ >7.24

The critical value for given value of α= 0.05 is 1.645.

The Test statistic is Z = (x̄ – μ) / (σ / √n)

Z=(8.40-7.24)/(2.066/SQRT(39))

Z=(1.16)/(2.066/6.244)

Z=1.16/0.3308

Z=3.506

**Module 5:**

**Application, table

Description automatically generated**

***Inference from the Regression Analysis***

R Square= 0.4651

Adjusted R Square=0.4593

P Value= 7.2943E-26

*Coefficients of Intercept=* 6.90263249

*Coefficients of Votes =* 0.000300107

*Coefficients of Viewership=* 0.06663101

From the R square value of 0.4651, we can infer that only 46.51 % of variation of a dependent variable is explained by the independent variable.

Adjusted R Square=0.4593, Adjusted R-square adds precision and reliability by considering the impact of additional independent variables that tend to skew the results of R-square.

As the difference between R Square and Adjusted R Square is very less, there is less to no additional independent variables that skew the results of R-square.

The Coefficients of both Votes and Viewership and Intercept are positive.

The P-Values of both Votes and Viewership are less than the Alpha given so they are both significant.