FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SECI2143-03 PROBABILITY & STATISTICAL DATA ANALYSIS

REPORT OF PROJECT 2

NO	NAME	MATRIC ID
1	ARIF AMIRUDDIN BIN SADIRAN	A19EC0022

LECTURER'S NAME: DR. ARYATI BINTI BAKRI

DATE OF SUBMISSION: 27 JUNE 2020

INDEX

Abstract— **The objective of this study is to investigate the relationship of 6 different types of alcohols and its consumption among the teenagers in Germany who are between 12 to 26 years old by performing and finding inferential statistics such as hypothesis testing, correlation and regression model between these two variables.**

1. INTRODUCTION

Alcohol which is an ethanol is the ingredient found in alcoholic beverages. Alcohol formed when sugar is breakdown without the present of oxygen. As sugar is available in different type of foods, each will create different type of alcoholic beverage.

Alcohol beverages come with its pros and cons, while alcohol could act as a stimulant, inducing feelings of euphoria and talkativeness, but in a large dose of alcohol at one session can lead to drowsiness, respiratory depression, coma or even death.

This study aims to investigate the different type of alcohol with each consumption among teenagers in Germany with ages between 12 to 26 years old.

2. METHODOLOGY

The data of this study is collected by social science research institutes, Institut für Jugendforschung, GfM-GETAS/WBA GmbH and forsa Gesellschaft für Sozialforschung und statistische Analysen mbH. The data is provided by German Ministry of Health Education. The data is obtained through a database website;

https://www.kaggle.com/fabiolabusch/alcohol-and-drug-consumption-of-german-teens

The data are selected from the dataset. The data will then be sampled for hypothesis testing to determine whether there is enough statistical proof to support the null hypothesis. The sample is normally distributed and plotted using RStudio.

3. RESULTS AND DISCUSSION

A. *Hypothesis testing*

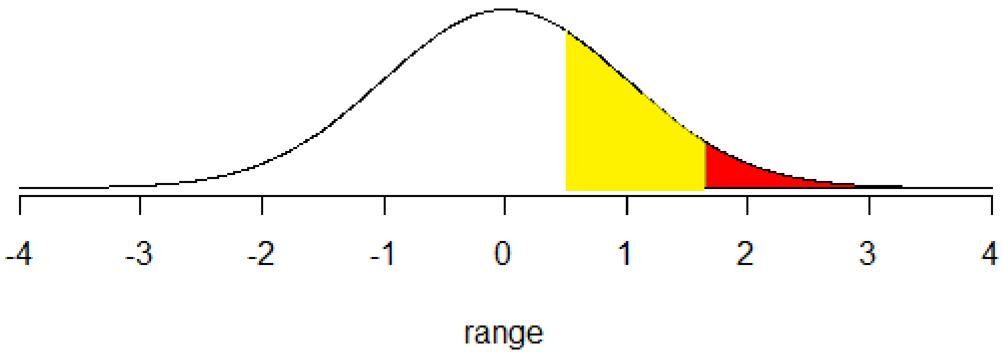


Fig 1.1: Bell Curve Graph

Based on the study “Alcohol consumption statistics” conduct by Euro Stat, the consumption of alcohol for once a month in Germany is 21 percent. Hence, the null hypothesis, H0 and alternative hypothesis, H1 is:

H_o: $\mu = 0.21$

H_1: $\mu > 0.21$

Where μ is the mean consumption of alcohol for atleast once a month.

The mean of collected data is 0.294. The standard deviation can be calculated by using formula:

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})2}{n-1}}$$

The standard deviation is 0.1053242.

A 95% level of confidence is used to test the claim of this study that the consumption of alcohol for once a month in Germany is 21 percent. The critical value for 0.05 confidence interval is 1.645. The test statistic of mean is calculated by using formula

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

and is equal to 0.675.

$\bar{x}$	μ	s	Test Statistic	Critical value
0.2940667	0.21	0.1053242	0.6747015	1.645

Since the test statistic < critical value. Hence, we fail to reject the null hypothesis, H_o as there is insufficient evidence to support the claim that the consumption of alcohol for once a month in Germany is 0.21. This show that the obtained from Euro Stat is still true.

B. *Correlation*

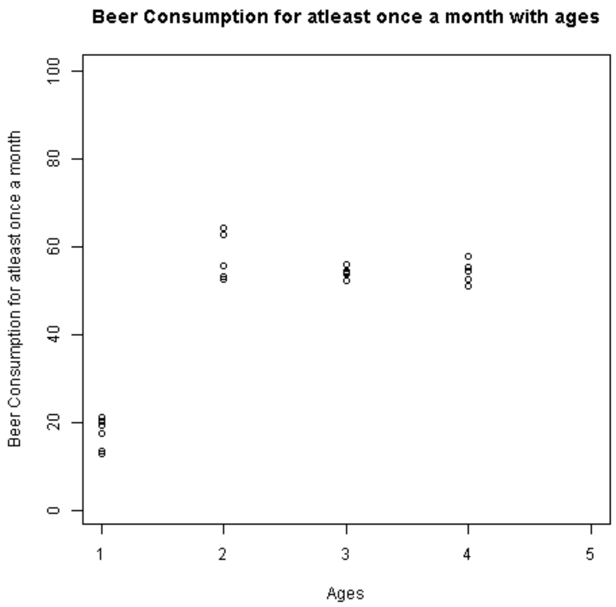


Fig 1.2: Scatter plot of beer consumption for atleast once a month against different ages.

Value Representation	1	2	3	4
Ages	12-16	16-18	18-22	22-26

In the correlation test, we had analysed the strength of the relationship between the beer consumption against different ages.

The coefficient correlation, r is calculated using the formula below.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The correlation coefficient is equal to 0.7234307. The results show that the relationship is moderately strong positive relation. Thus, as ages increased, the beer consumption for atleast once a month will also increase. A 0.05 level of significance is used to test the linear relationship between the two variables. H_o is assumed as the beer consumption for atleast once a month against ages has no linear correlation relationship.

H_o: $\rho = 0$; p = population correlation coefficient

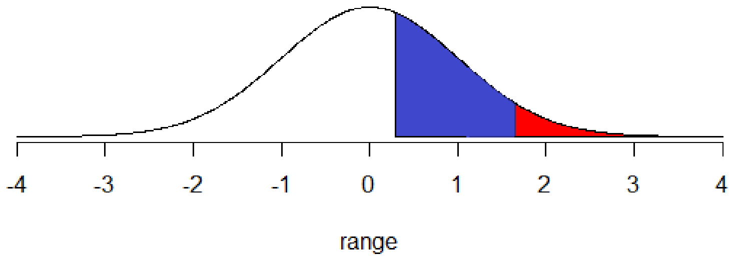
H_1: $\rho \neq 0$

The test statistics formula as below.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The sample size n is equal to 22 with correlation r is equal to 0.7234307. The test statistic is equal 30.3549 while the critical value is equal to 0.36.

Test Statistic, t	α	Degree of Freedom, v	Critical Value
30.3549	0.05	20	0.36



Since the test statistic is larger than critical value (30.3549>0.36), H_o at α = 0.05 is rejected. There is sufficient evidence to conclude that there is a linear relationship between the beer consumption for atleast once a month against ages.

C. Linear Regression

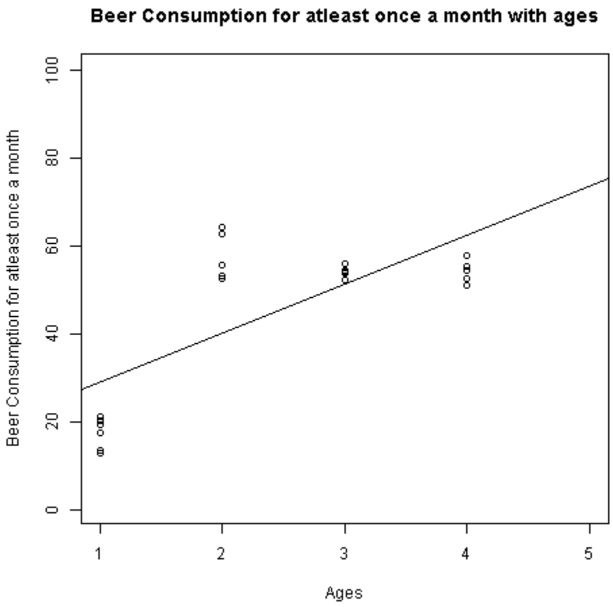


Fig 1.3: Scatter plot of beer consumption against ages with the regression line.

Value Representation	1	2	3	4
Ages	12-16	16-18	18-22	22-26

Linear regression is a method of predictive modelling in order to find the relationship between 2 continuous variable, variable X and variable Y. However, it could only be used to predict the value of variable Y when the value of variable X is known.

The general equation is as follows:

$Y = \beta 0 + \beta 1x + \varepsilon$

H_o: $\beta 1 = 0$

H_1: $\beta 1 \neq 0$

The regression shows the result of the relationship between beer consumption for atleast once a month against ages with dependent variable is beer consumption for atleast once a month and Independent variable as ages. The estimated regression model is obtained with

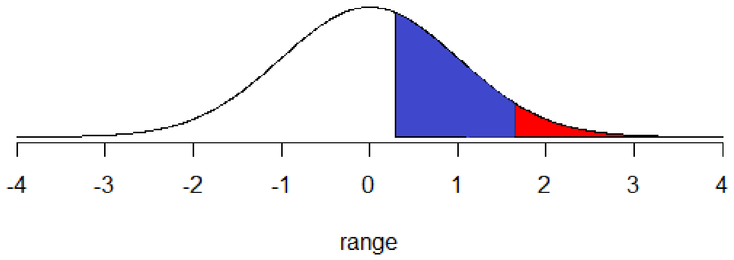
$Y = 18.239 + 11.123x$

β0 which is equal to 18.239 is equal as the estimated average value of Y when the value of X is equal to zero, while β1which is equal to 11.123 measures the estimated change in the average value of Y as a result of a one-unit change in X.

R2 = SSR/SST

The value of coefficient of determinant, R2 is equal to 0.5233086. Thus, it could be considered as weaker linear relationship between beer consumption for atleast once a month and ages as some but not all of the variation in variable Y is explained by variation in X.

Test Statistic, t	α	Coefficient of Determinant	Critical Value
4.320207	0.05	0.5233086	0.36



Since the test statistic is larger than critical value (4.320207>0.36), H_o at α = 0.05 is rejected. There is sufficient evidence to conclude that the age effect the beer consumption for atleast once a month.

D. ANOVA test

A one-way ANOVA test method is used to compare 6 different types of alcohol respectively with its consumption for atleast once a month. Significance level of 0.05 is used to test data. The null hypothesis, H0 and alternative hypothesis, H1 is:

H_0: The mean of consumption for 6 different types of alcohol is equal.
H_1: At least one mean of consumption for 6 different types of alcohol is different.

The results are shown below.

Type of Alcohol	Beer	Wine	Hard Liquor	Alcopops	Alcopops Beer	Alcopops Liquor
Consumption of alcohol for atleast once a month	43.0	36.0	23.1	23.1	23.1	23.1
	43.3	31.7	25.2	38.4	22.7	19.0
	47.3	29.9	25.2	39.8	26.1	11.7
	44.6	25.5	20.5	33.2	32.4	6.3
	44.8	28.9	23.9	37.0	30.9	5.7

Fig 1.6: Description table

Type of Alcohol	Mean	Standard Deviation	Variance Between Sample	Variance Within Sample	F	Critical Value	Degree of Freedom Numerator	Degree of Freedom Denominator
Beer	44.6	1.70147	51.21412	13.54854	3.780048	2.621	5	24
Wine	30.4	3.858756						
Hard Liquor	23.58	1.940876						
Alcopops	37.82	2.939728						
Alcopops Beer	27.3	4.182702						
Alcopops Liquor	15.8	11.60733						

Fig 1.7: ANOVA table

Since the test statistic F is larger than critical value (3.780048>2.621), H_o at $\alpha = 0.05$ is rejected. There is sufficient evidence to conclude that the mean consumption of alcohol is different.

4. Conclusion

Based on the hypothesis testing, we fail to reject the null hypothesis, Ho as there is insufficient evidence to support the claim that the consumption of alcohol for once a month in Germany is 0.21. This show that the obtained from Euro Stat is still true. Next, in correlation test we reject the H_o. This shows that we have sufficient evidence to conclude that there is a linear relationship between the beer consumption for atleast once a month against ages.

The estimated regression model is able to be calculated $Y = 18.239 + 11.123x$, and this equation could be used to predict the beer consumption for different ages. In Anova, there is sufficient evidence to conclude that the mean consumption of different type of alcohol is different.

4. Acknowledgment

Sincere appreciation is extended to our lecturer, Dr Aryati binti Bakri for helping on this study.

5. References

[1] https://ec.europa.eu/eurostat/statistics-explained/index.php/Alcohol_consumption_statistics

[2] https://www.alcohol.org.nz/alcohol-its-effects/about-alcohol/what-is-alcohol

[3] https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics