UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143-04 KEBARANGKALIAN STATISTIK & ANALISIS DATA

PROJECT 2

SECTION : 04 – 1SECR

COURSE NAME : BACHELOR OF COMPUTER SCIENCE – COMPUTER NETWORKS & SECURITY

NAME: MUHAMMAD ISYRAFF IRFAN BIN MOHAMMAD RIZAL

LECTURER'S NAME: Dr. SUHAILA BINTI MOHAMAD YUSOF

DATE OF SUBMISSION: 27/6/2020

Table of content

1. **Introduction**

Suicide is a topic that unlikely to talked about often but suicide does not only affect adults, but it also affects teenagers as well. Suicide is now among the three leading causes of death among those aged 15-44 (male and female). Suicide attempts are up to 20 times more frequent than completed suicides. Mental health disorders such as depression and substance abuse are associated with more than 90% of all cases of suicide. However, suicide results from many complex factors and it is more likely to occur during periods of socioeconomic, family and individual crisis such as loss of a loved one, unemployment and disassociation from one's community or other social/belief group and honor. Based on The World Health Organization (WHO) estimates that each year approximately one million people die from suicide, which represents a global mortality rate of 16 people per 100,000 or one death every 40 seconds.

Suicide doesn't care about your gender, age and where you come from. It can happen to anyone. Genders differences in suicide rates have been shown to be significant. The purposes in this report is to study the suicide rate between gender in every country to see if there are differences of suicide rates between two gender which is male and female.

2. Hypothesis testing

For the hypothesis testing, I tested on two sample to compare the mean of number of suicides per 100,000 population happened between male and females. For my null statement, the mean of number of suicides for male are equal with the mean of suicides rate for female which is equal to 0. For my alternative statement the mean of the mean for male and female are not equal to 0 meaning that the mean for both genders are different.

μ_1= mean of suicides rate per 100,000 population for male in every country.

μ_2= mean of suicides rate per 100,000 population for female in every country.

$$H_0 = \mu_1 = \mu_2$$

$$H_1 = \mu_1 \neq \mu_2$$

To find if either we reject or fail to reject the null statement, I calculate the t-test in R.

```
        Welch Two Sample t-test

data:  suicide_per_100k pop by gender
t = -70.961, df = 16599, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.25656 -14.43637
sample estimates:
mean in group female   mean in group male
            5.392866            20.239329
```

Figure 1 T-test

In the output given in R, we can see the mean of suicide is 5.3929 and 20.2393 for female and male respectively. I used the data from suicide_per_100k pop by gender since I want to find the sample mean between male and female. The suicides_per_100k pop is equal to suicides_no divided by 100,000 of population. The reason why I used suicides_per_100k pop to find mean_suicides is because there might be an ambiguity since male population is higher than female population.

Based on the result in R, the p-value is 2.2x10-16 and the significance level, $\alpha = 0.05$, we reject null hypothesis that the mean for male and female is equal to zero. There is sufficient evidence that the mean of suicide rate between male and female are not equal to zero.

```
> qt(0.025,19)
[1] -2.093024
> -qt(0.025, 19)
[1] 2.093024
```

Figure 2 Critical value

The qt in R is the critical value which is 2.0930 and -2.0930. The t-test value I got in R is t= -70.961 which is lesser than the critical value, -2.0930. This also shows that we reject null hypothesis.

The sample mean, standard deviation and variance for male and female:

	gender	mean_suicides	sd_suicides	variance_suicides	sum_pop
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	female	5.39	7.36	54.2	26272781857
2	male	20.2	23.6	555.	25049376579

The mean_suicides is the mean for suicides per 100k population.

Conclusion: Since p-value = $2.2 \times 10^{-16} < \alpha = 0.05$, we reject null hypothesis. There is sufficient evidence that the mean of suicide rate for male and female are not equal.

3. Correlation – Pearson's Technique

```
#correlation between population and suicides number
x = master$suicides_no
y = master$`suicide_per_100k pop`
plot(x, y, main = 'Correlation between suicides number and suicides per 100k population',
     xlab = 'suicides number', ylab = 'suicides per 100k population',cex.axis=0.9)
abline(lm(y~x))
cor.test(x, y)
```

Figure 3 Code for correlation

The correlation in this report is to study the relationship between the number of suicides and the number of suicides per 100,000 population. In R, the scatter diagram will be plotted and cor.test function is used to calculate the correlation between the number of suicides and the suicides per 100,000 population.

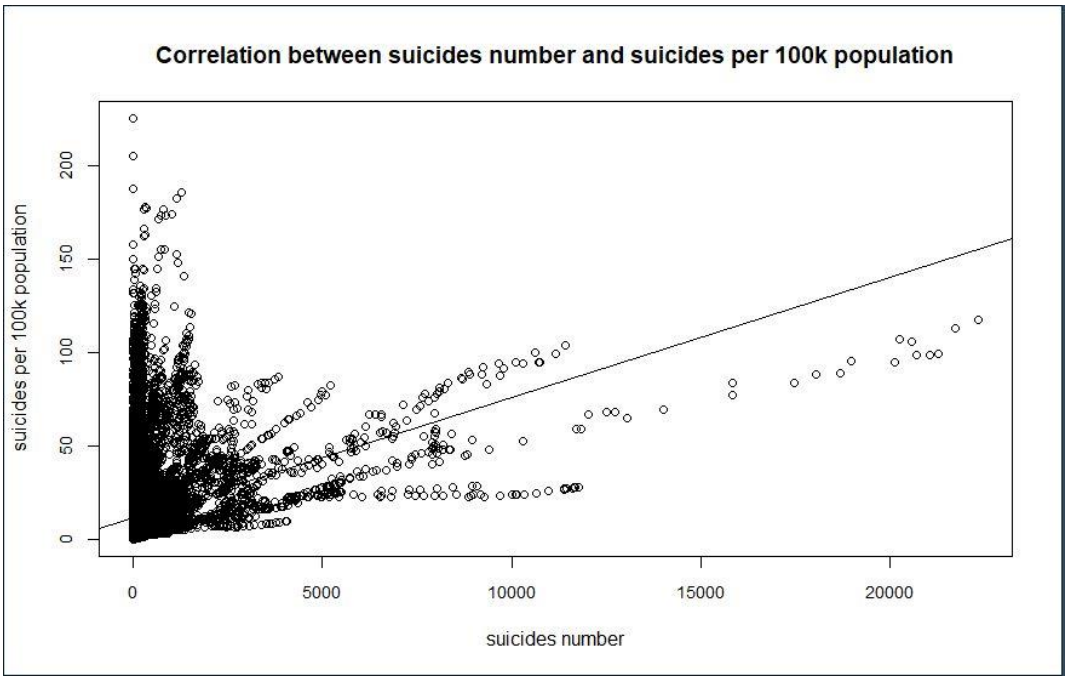

Figure 4 Correlation graph

The value I get from R,

```
        Pearson's product-moment correlation

data:  x and y
t = 53.725, df = 27818, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2959197 0.3172125
sample estimates:
      cor
0.3066045
```

Figure 5 Result for correlation

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

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

H0: $\rho = 0$ $r = 0.3066045$ $t = 53.725$

H1: $\rho \neq 0$ $\alpha = 0.05$ *P-value* $= 2.2 \times 10^{-16}$

Statically, the correlation coefficient *r* measures the strength and direction of a linear relationship between two variables using scatterplot. The value of *r* is between +1 and -1. Based on R, the value of *r* is 0.3066045 which is near to zero. This shows that the relationship between suicides number and number of suicides per 100k population have a weak relationship. It also indicates that these two variables have correlation since the p-value = 2.2×10^{-16} is less than significant level, $\alpha = 0.05$.

Conclusion: Since the p-value = $2.2 \times 10^{-16} < \alpha = 0.05$, we reject null hypothesis. There is sufficient that there is correlation between the suicides number and number of suicides per 100,000 population.

4. Regression

```
x = master$year
y = master$`suicide_per_100k pop`
model = lm(y~x)
summary(model)
plot(x, y, main = 'Regression between year and suicides per 100k population', xlab = 'year', ylab = 'suicides per 100k
abline(model)
```

Figure 6 Code for Regression

The regression in this study is about the relationship between year and the suicides per 100,000 population. By using R, the data of years is stored in x variable whereas the suicides per 100,000 population is stored in y variable. The summary is used to produce the result of the regression model. Then, the graph is plotted with abline that can add regression line.

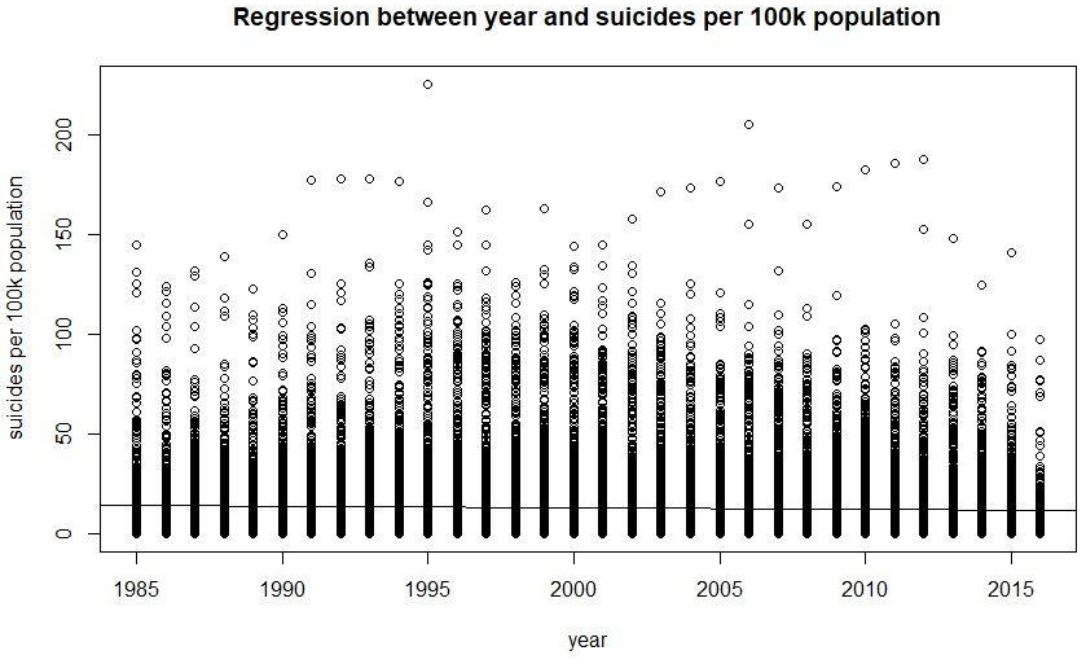

Figure 7 Regression graph

The value for regression calculated in R,

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-14.237 -11.682  -6.867   3.802 211.607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 187.72638   26.84423   6.993 2.75e-12 ***
x            -0.08740    0.01341  -6.516 7.35e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.95 on 27818 degrees of freedom
Multiple R-squared:  0.001524,  Adjusted R-squared:  0.001488
F-statistic: 42.46 on 1 and 27818 DF,  p-value: 7.353e-11
```

Figure 8 Result for Regression

H0: $\beta 1 = 0$ P-value: Intercept = 2.75×10^{-12} $\hat{y}_i = b_0 + b_1 x$

H1: $\beta 1 \neq 0$ P-value: Slope = 7.353×10^{-11} $\hat{y}i = 187.72638 + (-0.087)\ x$

The slope's coefficient estimate value shows that as year increases, the suicides per 100k population decreases by 0.08740. The small p-values of intercept, 2.75 x10-12 and slope, 7.353×10^{-11} indicates that we reject null hypothesis which means there is relationship between the suicides per 100,000 population and the year.

Conclusion: Since both p values are less than the significant level $\alpha = 0.05$, we reject null hypothesis. There is sufficient evidence at 95% confidence level that the year affect the suicides per 100k population.

5. Chi-Square (Goodness of fit test)

```
#chi square test
x = master$`suicide_per_100k pop`
prod = chisq.test(x, correct = FALSE)
str(prod)
prod
```

Figure 9 Code for chi-square

The goodness of fit test in this study is about chi-square test for the suicide per 100k population in every country. From R, the total of suicides per 100,000 population stored in x variable. The chi-square function used in R will returns the numbers from chi-square distribution in this coding.

```
        Chi-squared test for given probabilities

data:  x
X-squared = 780426, df = 27819, p-value < 2.2e-16
```

Figure 10 Result for chi-square

H_0= Proportion of number of suicides rate per 100,000 population in every country is the same.

H_1= At least one of the proportions is different from the others.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

X^2 = 780426, df= 27819, p-value =2.2×10^{-16}

The large value of X^2 shows that the data does not fit very well since only small chi square will fit to be the expected data. Since the p-value, 2.2×10^{-16} is smaller than $\alpha = 0.05$, the null hypothesis is rejected. There is insufficient evidence to conclude that the proportion of number of suicides rate per 100,1000 population in every country is the same.

6. ANOVA

```
> ANOVA1=aov(suicides_no~country)
> ANOVA1
Call:
   aov(formula = suicides_no ~ country)

Terms:
                     country   Residuals
Sum of Squares    9222936581 13413117888
Deg. of Freedom          100       27719
```

Figure 11 ANOVA result

The one- way analysis of variance (ANOVA) is an extension of independent two sample t-test for comparing means in a situation where there are more than two groups or types. In my analysis, I tested the suicides number separated by in every country.

```
> summary(ANOVA1)
              Df    Sum Sq  Mean Sq F value Pr(>F)
country      100 9.223e+09 92229366   190.6 <2e-16 ***
Residuals  27719 1.341e+10   483896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12 Summary for ANOVA

H_0= The mean for the number of suicides is the same for every country

H_1= At least one mean is different.

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{ns_{\bar{x}}^2}{s_p^2}$$

F value= 190.6 p value= $2x10^{-16}$ $\alpha = 0.05$

Since the p value is lower than the significance level, $\alpha = 0.05$, we can reject the null hypothesis. There is sufficient evidence that the number of suicides has at least one mean that is different in every country.

From the table of Tukey multiple comparisons of means 95% family-wise confidence level, by looking at the "diff" and "p adj" columns, we can see which month has the significant difference in number of suicides between every country because the p-values is lower than the significance level, $\alpha = 0.05$. By plotting TUKEYHSD(ANOVA1) in R, I can visualize and analysis the significant difference. From the graph below, we can see that there is a significant difference because not all line for every pair crossing the zero value.

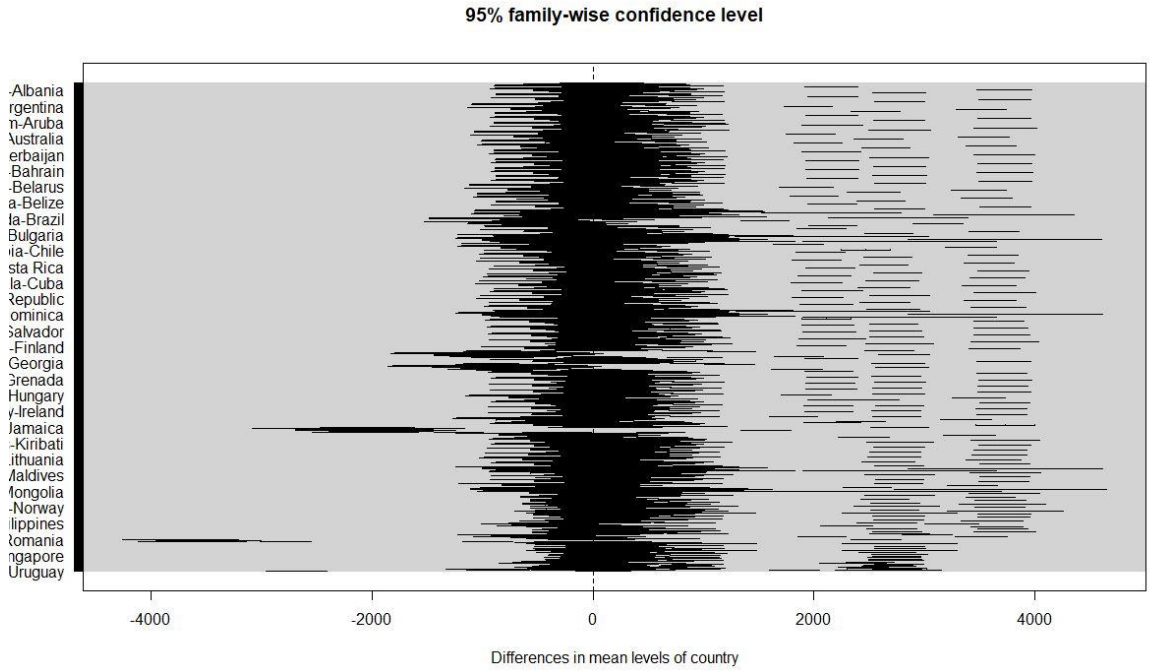

Figure 13 Tukey Test

7. Discussion

Based on the studies and analysis that I did earlier, the mean for the number of suicides per 100,000 population for male and female is 20.239329 and 5.392866 respectively. Using hypothesis testing for two sample, the differences of mean is tested and it is found that they have different mean and it shows that the number of suicides per 100,000 population for male is higher than female.

For the correlation using Pearson's technique, I test if there is a relationship between the number of suicides and the number of suicides per 100,000 population. As a result, the value I obtained for R is 0.3066045. Since the value of r is near to 0, it indicates that the relationship between the number of suicides and the number of suicides per 100,000 population are weak linear relationship. Therefore, we can conclude there is a slight relationship between those two variables. Hence, it shows the low number of suicides rate is related with low number of suicides per 100,000 population.

Furthermore, regression analysis is used to predict the value of a dependent variable based on the value of at least on independent variable. In this analysis, I tested two variable which is the number of suicides per 100,000 population with year. As a result, I found that there is a relationship between the number of suicides per 100,000 population and year. Thus, I can conclude that the number of suicides per 100,000 population also depends on year.

Moreover, I tested that if the proportion of confirmed deaths in each country is the same using chi-square goodness of fit test. As a result, the proportion of confirmed deaths is different for each country. In developing country, the number of suicides is higher in that area instead of developed country.

Lastly, I conducted analysis of variance (ANOVA) to compare the mean number of suicides separated by country. As a result, the mean number of suicides at least one of the countries has different mean. It shows that the number of suicides is different in every country depending on other factor such as the economics of certain country especially between high income country and low-income country.

8. Conclusion

After all the finding that I get from the statistical analysis test, I can conclude the suicides rate for male is higher than female. The difference role and societal expectation might be the factor of the differences in suicides rate where men are higher than female. The gender stereotype of men being strong and shouldn't admit they're struggling made them bottle up their emotion is one of the factors of higher suicides rate than female. As the year increases, there is minor decrease for the suicides rate per 100k population as the year increases. We can see that as the country develop and as the year increases where suicides awareness is increasing and people acknowledging it, it is one of the reasons why the suicides rate every year is decreasing.

Suicide is an occurrence that is preventable. It is important for teenagers to have knowledge about in how to deal with suicide so they can help their peer from committing suicide and at the same time they can detect symptom if they are having suicidal thoughts and go to seek from the professional. This is why suicide awareness and prevention are very important especially in developing country so they can have at least the knowledge and the awareness about suicide.

9. References

1. WHO. (2019). *Suicide: one persone dies every 40 seconds.*
2. WHO. (2016). *Suicide rates per 100 000 population.*
3. World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/