UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

INDIVIDUAL CASE STUDY REPORT:

PROJECT 2

SECI2143 – PROBABILITY AND STATISTICAL DATA ANALYSIS

SECTION : 04 - 1 / SECR

COURSE : BACHELOR OF COMPUTER SCIENCE (COMPUTER NETWORK AND SECURITY)

NO	NAME	STUDENT ID
1	MUHAMAD NAZREEN BIN MUBIN	A19EC0104

LECTURE'S NAME : DR SUHAILA BINTI MOHAMAD YUSUF

DATE OF SUBMISSION : 27TH JUNE 2020

INTRODUCTION

I have been given a handout of the instructions for this assessment via e-learning on May 17. Based on the instructions given, I was tasked to inspect about the inferential statistics on the following website, https://www.who.int/healthinfo/statistics/data/en/ using my understanding of the topics that have been learned. The case study discussed about the suitable inferential statistics based on the data found in the website. The hypothesis testing included was hypothesis testing 1-sample, correlation, regression and ANOVA test.

Secondary data is a data that was already collected through primary sources. The data was collected by the researchers for a particular research or project, and by then make it available for all to make use the data for their own interest. Thus, it can be said that a data defined as secondary may be seen as primary sources by another researchers. Because of the advent of the internet and modern devices, the access to secondary data had became much easier. The sources that can be searched via websites and blogs records that do not existed in the past. Furthermore, we can access to traditional ways of collecting data such as books, newspaper, journal, government records and diaries via the internet.

The data found was for measuring maternal mortality accurately is difficult except where comprehensive registration of deaths and of causes of death exists. Mortality rate or death rate is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time. World Health Organization (WHO) is a specialized agency of the United Nations responsible for international public health.

HYPOTHESIS TESTING

A. Hypothesis Testing 2 – Sample

Both genders are being analysed to determine how they affect the mean life expectancy at birth. At **α=0.05** level, a test is conducted on 192 people to test the claim that the mean life expectancy at birth are differ based on the genders. Here, the hypotheses are …

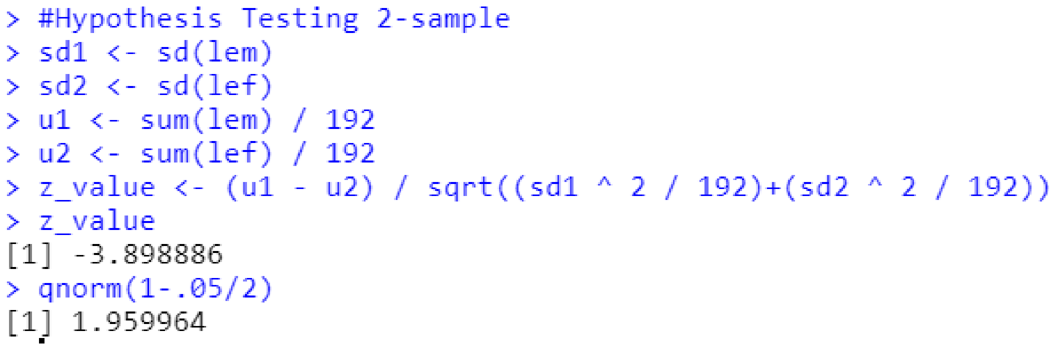
H0: μ1 = μ2

H1: μ1 ≠ μ2

where,

μ1 = mean life expectancy at birth of males

μ2 = mean life expectancy at birth of females

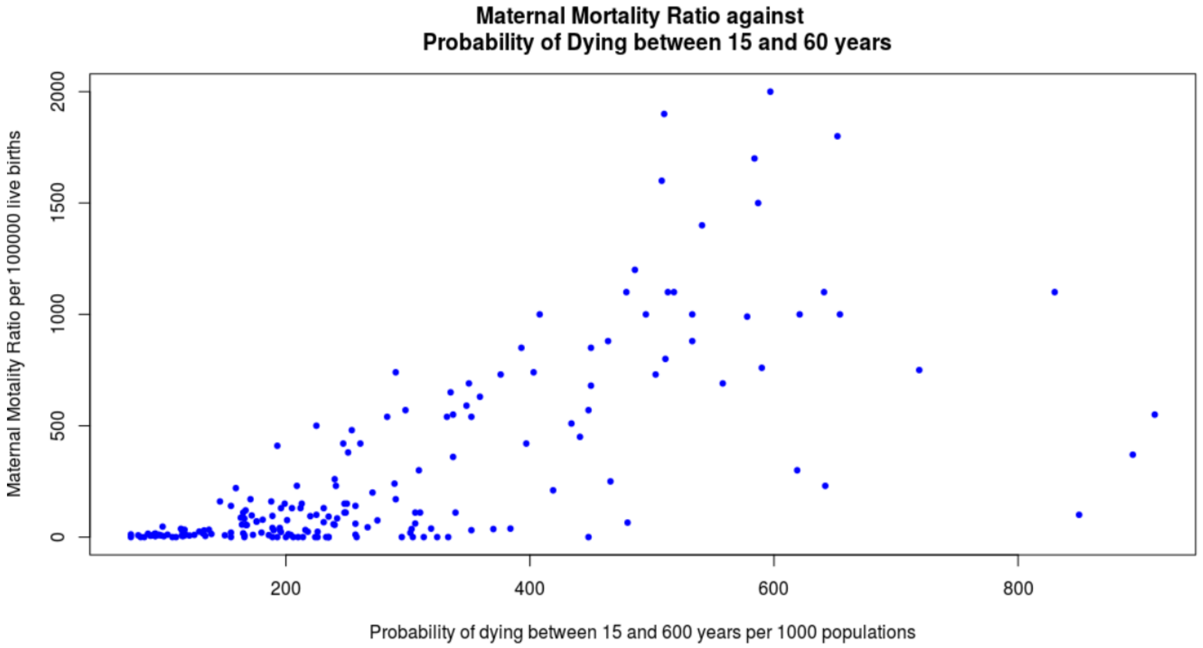


```
> #Hypothesis Testing 2-sample
> sd1 <- sd(lem)
> sd2 <- sd(lef)
> u1 <- sum(lem) / 192
> u2 <- sum(lef) / 192
> z_value <- (u1 - u2) / sqrt((sd1 ^ 2 / 192)+(sd2 ^ 2 / 192))
> z_value
[1] -3.898886
> qnorm(1-.05/2)
[1] 1.959964
```

The figure above shown the calculation of z-statistical value and z-critical value using RStudio. The data **z_value** is the statistical value calculated in the hypothesis testing while the line 'qnorm(1-.05/2)' is the critical value calculated using the function found through a website, https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R-Manual/R-Manual10.html.

The decision to be made from the values researched is to reject the null hypothesis, H0 since the statistical value (3.8989) is more than the critical value (1.96). Hence, there is sufficient evidence to claim that the mean life expectancy at birth are differ based on the genders.

B. Correlation



The figure above shown the scatter plot of maternal mortality ratio against probability of dying between 15 and 60 years. The plotted data seem to have a linear relationship between the two variables. By the look of it, both variables can form a positive relation. To show the relationship between these two variables, the coefficient correlation, r is calculated using RStudio. These two variables are used to calculate the coefficient correlation using the following formula below.

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

where,

x = probability of dying between 15 and 60 years for males

y = maternal mortality ratio

The correlation coefficient is 0.7132, a strong positives relation that proved the claim about the scatter plot. The correlation test is held to investigate the relationship between probability of dying per 1000 population in the range of 15 and 60 years of males and maternal mortality ratio. At **α=0.05** level, a test to claim that there is a relation between the probability of dying between 15 and 60 years for males and maternal mortality ratio. Here, the hypotheses are …

H0: ρ = 0

H1: ρ ≠ 0

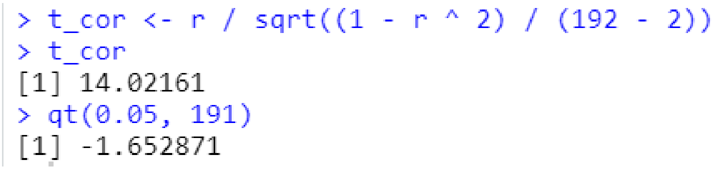
where,

ρ = population coefficient correlation

Using 0.05 significance level, the calculation for the test is done using the following formula.

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

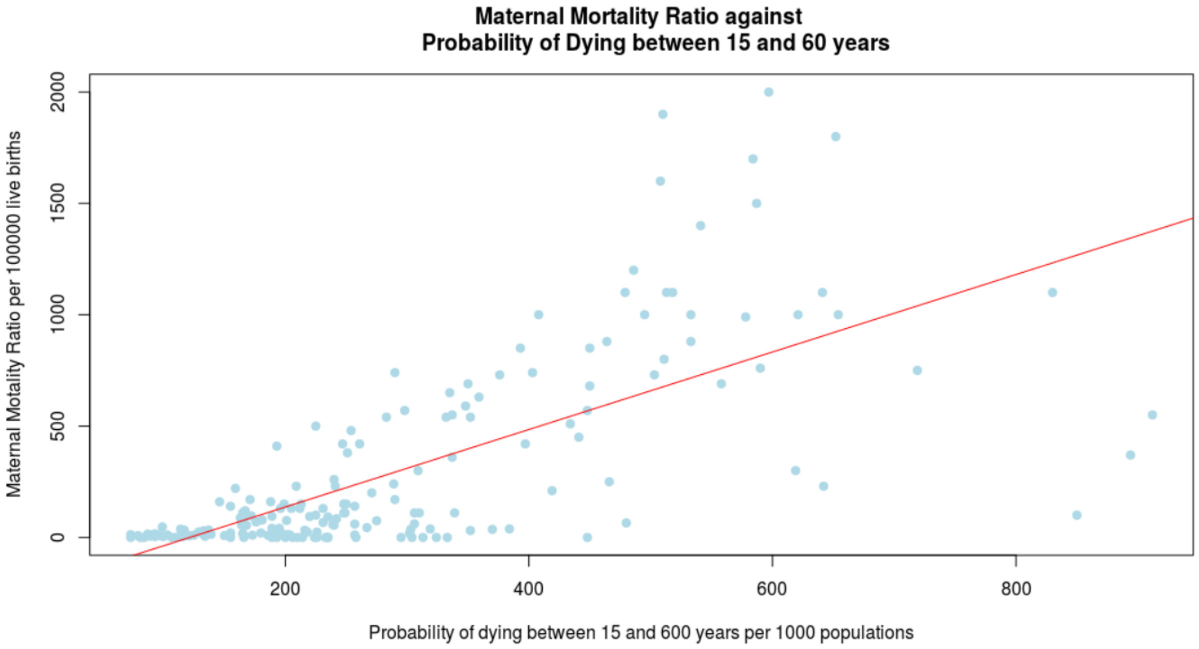
The formula was reproduced using RStudio and the result(t_cor) is the t statistical value as shown in diagram below.



```
> t_cor <- r / sqrt((1 - r ^ 2) / (192 - 2))
> t_cor
[1] 14.02161
> qt(0.05, 191)
[1] -1.652871
```

The decision based on the values calculated is to reject null hypothesis since the statistical value (14.022) is more than the critical value (-1.652). Therefore, there is enough evidence to conclude that a relation is exist between probability of dying per 1000 population in the range of 15 and 60 years of males and maternal mortality ratio al 0.05 significance level.

C. Regression



The scatter plot in the diagram above shown the maternal mortality ratio against probability of dying between 15 and 60 years. The red line in the graph is the line with equation calculated using the particular equation,

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

where,

y = dependent variable,

$\beta 0$ = the population intercept y,

$\beta 1$ = the population slope coefficient

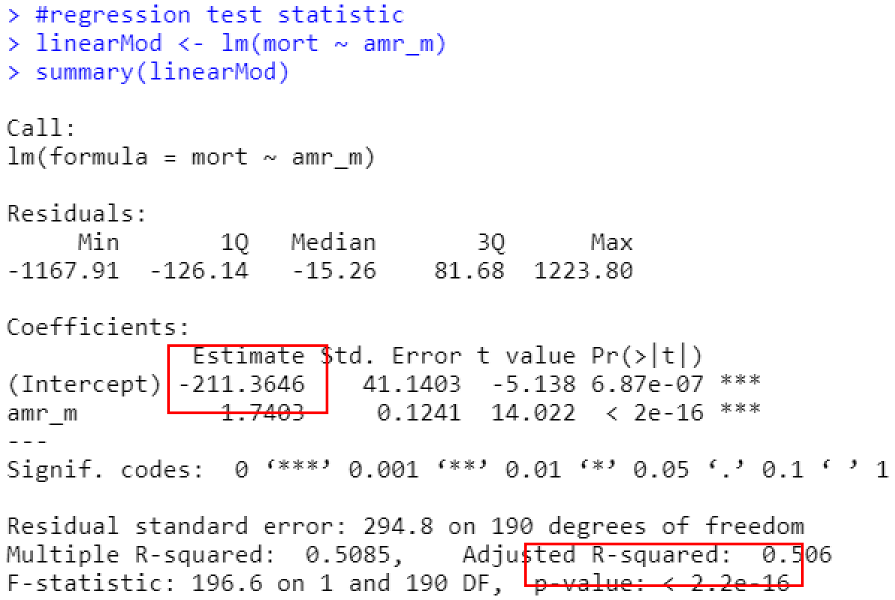
x = independent variable,

ε = random error component.

Based on the graph, the dependent variable will be the maternal mortality ratio while the independent variable is the probability of dying between 15 and 60 years. A regression test is conducted to inspect the relation between two variables mentioned. Here, the null and alternative hypotheses:

H0: $\beta 1 = 0$

H1: $\beta 1 \neq 0$



```
> #regression test statistic
> linearMod <- lm(mort ~ amr_m)
> summary(linearMod)

Call:
lm(formula = mort ~ amr_m)

Residuals:
     Min       1Q   Median       3Q      Max 
-1167.91  -126.14   -15.26    81.68  1223.80 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -211.3646    41.1403  -5.138 6.87e-07 ***
amr_m          1.7403     0.1241  14.022  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 294.8 on 190 degrees of freedom
Multiple R-squared:  0.5085,    Adjusted R-squared:  0.506 
F-statistic: 196.6 on 1 and 190 DF,  p-value: < 2.2e-16
```

The estimated regression calculated using RStudio, obtained $\hat{y} = -211.3646 + 1.7403x$.

The p-value is found to be less than 0.05, so the null hypothesis is accepted. According to the decision, there is insufficient evidence to conclude that there is no relation exist between the maternal mortality ratio against probability of dying between 15 and 60 years at 0.05 significance level.

D. ANOVA test

Using one-way ANOVA with equal sample sizes test, a comparison of the life expectancy at birth between the regions. Significance level of 0.05 is used to test data. Here, the hypotheses are …

$$H_0:\ \mu1 = \mu2 = \mu3 = \mu4 = \mu5 = \mu6$$

H_1: At least one is different.

where,

$\mu1$ = the mean life expectancy at birth in African region

$\mu2$ = the mean life expectancy at birth in region of the Americas

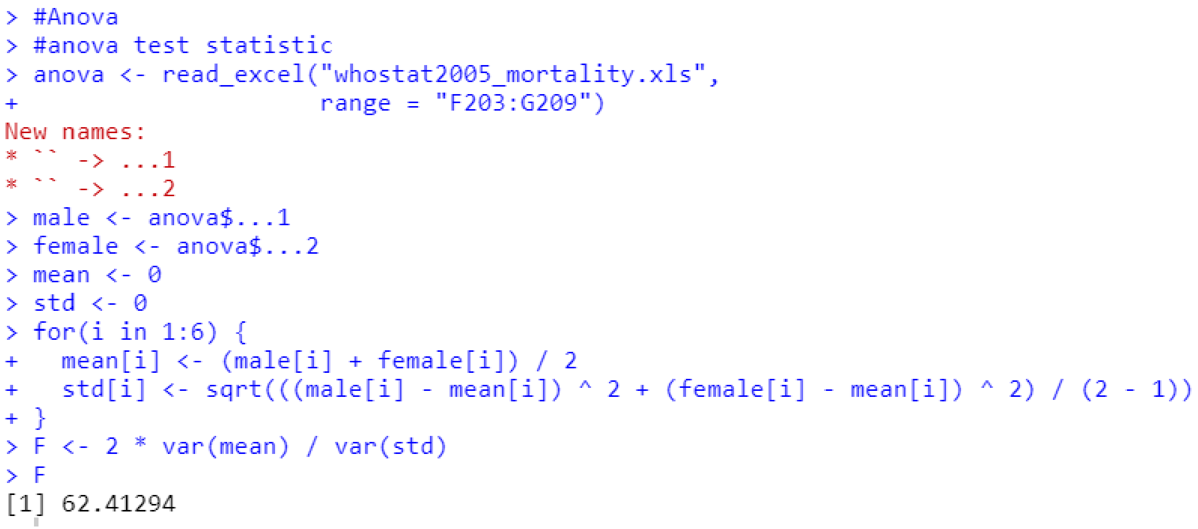
$\mu3$ = the mean life expectancy at birth in South-East Asia region

$\mu4$ = the mean life expectancy at birth in European region

$\mu5$ = the mean life expectancy at birth in Eastern Mediterranean region

$\mu6$ = the mean life expectancy at birth in Western Pacific region

The mean and standard deviation for each category is calculated using formulas. The formulas are applied in RStudio to calculate the test statistical value.



```
> #Anova
> #anova test statistic
> anova <- read_excel("whostat2005_mortality.xls",
+                     range = "F203:G209")
New names:
* `` -> ...1
* `` -> ...2
> male <- anova$...1
> female <- anova$...2
> mean <- 0
> std <- 0
> for(i in 1:6) {
+   mean[i] <- (male[i] + female[i]) / 2
+   std[i] <- sqrt(((male[i] - mean[i]) ^ 2 + (female[i] - mean[i]) ^ 2) / (2 - 1))
+ }
> F <- 2 * var(mean) / var(std)
> F
[1] 62.41294
```

The test statistic value of F is 62.41294. Since the test statistical value is more than the critical value, so the null hypothesis is rejected. Hence, there is enough evidence to conclude that the life expectancy is differ between the regions.

DISCUSSION

The dataset has collected about 192 data with 8 different variables. In this article, we studied about Hypothesis testing in R. We learned about the basics of the null hypothesis as well as alternative hypothesis. We read about T-test and μ-test. Then, we implemented these statistical methods in R. The objectives of RStudio is to define a function that takes arguments, return a value from a function, test a function, set default values for function arguments and also to explain why we should divide programs into small, single-purpose functions.

CONCLUSION

For the hypothesis testing on 2-sample, there is sufficient evidence to claim that the mean life expectancy at birth are differ based on the genders.

For the correlation test. there is sufficient evidence to conclude that a relation is exist between probability of dying per 1000 population in the range of 15 and 60 years of males and maternal mortality ratio al 0.05 significance level.

For the regression test, there is insufficient evidence to conclude that there is no relation exist between the maternal mortality ratio against probability of dying between 15 and 60 years at 0.05 significance level.

For the one way ANOVA test, there is sufficient evidence to conclude that the life expectancy is differ between the regions.

REFERENCES

1. https://stackoverflow.com/questions/5824173/replace-a-value-in-a-data-frame-based-on-a-conditional-if-statement
2. https://www.datacamp.com/community/tutorials/tutorial-on-loops-in-r?utm A7INWH5hFeuUvv3ikqgaAnYBEALw_wcB
3. http://r-statistics.co/Linear-Regression.html
4. https://cosmosweb.champlain.edu/people/stevens/WebTech/R/Chapter-9-R.pdf
5. https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/