UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143-04 PROBABILITY & STATISTICAL DATA ANALYSIS

Project 2

SECTION : 04 – 1SECR

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

NO.	NAME	STUDENT ID
1	AHMAD RAMADHAN SYUKRI BIN JAMALUDIN	A19EC0009

LECTURER'S NAME : Dr. Suhaila Binti Mohamad Yusuf

DATE OF SUBMISSION: 27 June 2020

Contents

Introduction

Data Description

The data that is used in this project is from Royal Malaysia Police. The data is about general road accident that has happened in Malaysia for the past years. The purpose of Royal Malaysia Police collecting this data is for annual report on road accident statistics. The data consist of year, registered vehicle, population, road crashes, road deaths, serious injury, slight injury, index per 10000 vehicles, index per 10000 population and index per billion vkt. The year is the only different type of data which is nominal and the others are categorize as ratio type of data. The data consist of year from 1997 to 2016.

Purpose of Study

The data that is used in this study is year, registered vehicles, road crashes and road deaths. The purpose of this study is to see whether the mean of death caused by road accidents is equal to 6000, is there any relationship between registered vehicle and road crashes, to see the impact of the number of car crash and the number road death and to see if the probability of having road death is equal in the 20 years of study.

To achieve these objectives, I used several test which are hypothhesis testing on 1 sample, Pearson's product-moment correlation, regression model and chi-square test one way contingency respectively.

Statistical Analysis

Hypothesis Testing for 1 Sample

For the first part, is hypothesis testing for 1 sample. The variable that is used from the data is road deaths. I will compare the mean of the number of death by car accident is wether equal to 6000 or not per year (with $\alpha = 0.05$) in 20 years. The coding that is used in Rstudio to complete this is as below:

```
#hypothesis-1 sample
t.test(main.dat$`Road Deaths`,mu = 6000, conf.level = 0.95)
qt(0.025,19)
-qt(0.025, 19)
```

The output of the program is as below:

```
> #hypothesis-1 sample
> t.test(main.dat$`Road Deaths`,mu = 6000, conf.level = 0.95)

        One Sample t-test

data:  main.dat$`Road Deaths`
t = 4.3344, df = 19, p-value = 0.0003573
alternative hypothesis: true mean is not equal to 6000
95 percent confidence interval:
 6214.06 6613.84
sample estimates:
mean of x
  6413.95

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

The H0 and H1 are:

H_0: μ = 6000

H_1: μ ≠ 6000

Data extraxt from output:

$t = 4.3344$ $\quad critical\ value: t_{(0.025,19)} = (-2.093024, 2.093024)$

$p - value = 0.0003573$

Since p-value < α (0.0002573 < 0.05) and t test statistic > t critical value (4.3344 > 2.093024), I choose to reject H0. The conclusion is, at 95% confidence interval, there is enough evidence to conclude that the mean of death cause by road accident in 20 years is not equal to 6000.

Correlation

For the next part, I use correlation to see wether there is a relaionship between registered vehicle and road crashes. The coding is as below:

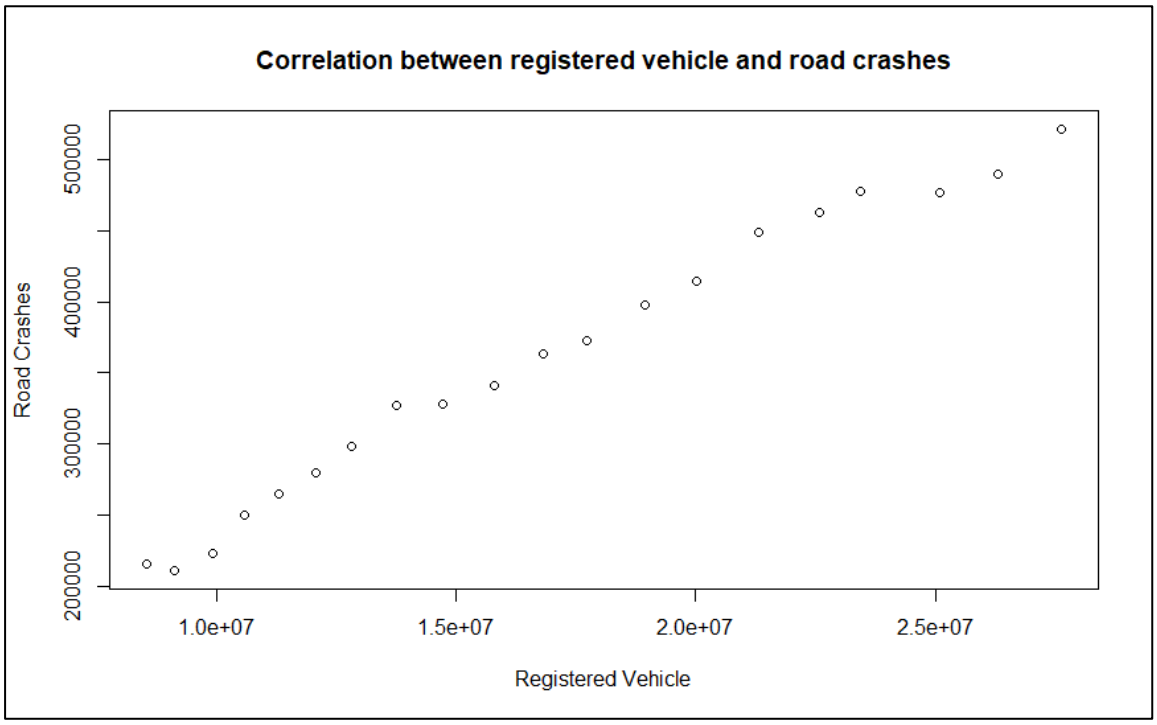
```
#correlation-between registered vehicle and road crashes
x = main.dat$`Registered Vehicles`
y = main.dat$`Road Crashes`
plot(x, y, main = 'Correlation between registered vehicle and road crashes', xlab = 'Registered Vehicle', ylab = 'Road
cor.test(x, y)
```

The output is as below:

```
> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = 33.213, df = 18, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9792758 0.9968773
sample estimates:
      cor
0.9919396
```



The H0 and H1 are:

H_0: $\rho = 0$ (no relationship)

H_1: $\rho \neq 0$ (has relationship)

Data extract from output:

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

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

$t = 33.213$

$p - value = 2.2 \times 10^{-16}$

$r = 0.9919396$

$\alpha = 0.05$

Since my p-value < α ($2.2 \times 10^{-16} < 0.05$), I choose to reject H0. At 95% confidence interval, there are sufficient evidence to conclude that there are relationship between registered vehicle and road crashes.

The r is to show how strong the relationship between the variables. The range of r is between -1 strongest negative relationship and +1 strongest positive relationship. In this case, the r between registered vehicle and road crashes is equal to 0.9919396 which means it has strong positive relationship.

Regression

For regression, I use to test if road crashes affect road deaths with $\alpha = 0.05$. The coding is as follow:

```
#regression-between road crashes and road deaths
x = main.dat$`Road Crashes`
y = main.dat$`Road Deaths`
model = lm(y~x)
summary(model)
plot(x, y, main = 'Regression between road crashes and road deaths', xlab = 'Road Crashes', ylab = 'Road Deaths')
abline(model)
```

The output is as below:

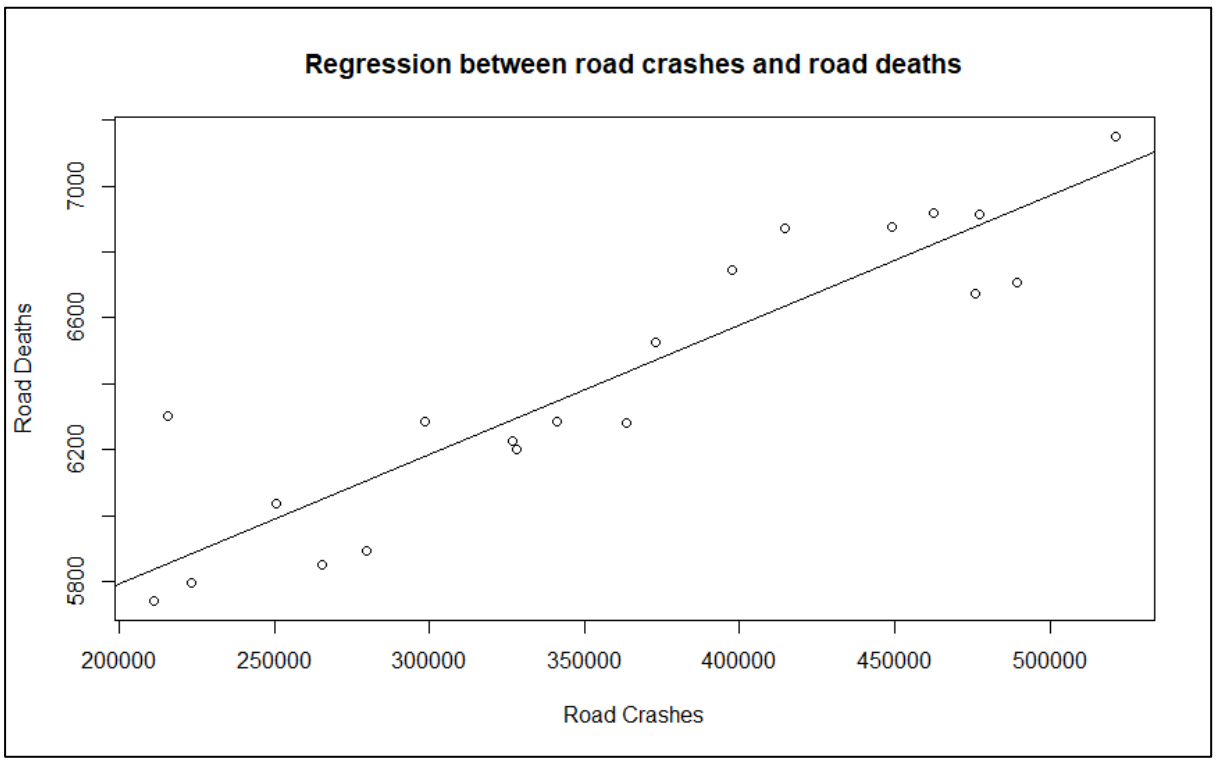
```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-223.08 -110.72  -12.96   99.90  447.02

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.010e+03  1.521e+02  32.930  < 2e-16 ***
x           3.921e-03  4.099e-04   9.564 1.77e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 177.9 on 18 degrees of freedom
Multiple R-squared:  0.8356,    Adjusted R-squared:  0.8264
F-statistic: 91.47 on 1 and 18 DF,  p-value: 1.765e-08
```



The H0 and H1 are:

$H_0: \beta_1 = 0$ (No linear relationship)

$H_1: \beta_1 \neq 0$ (Linear relationship does exist)

Data extraxt from output:

$\hat{y}_i = b_0 + b_1 x$

$b_0 = 5.010 \times 10^{03}$

$b_1 = 3.921 \times 10^{-03}$

$\hat{y}_i = 5.010 \times 10^{03} + 3.921 \times 10^{-03} x$

$p - value: Intercept = 2 \times 10^{-16}$

$p - value: Slope = 1.77 \times 10^{-08}$

Since both p-value is less than α = 0.05, I reject the H0. As for the conclusion, at 95% confidence interval, there is sufficient evidence to conclude that road crashes affect road deaths. We can also predict any variable from the equation that we obtain from the calculation.

Chi-Square Test One Way Contingency

For the next test, I used the chi-square test to determine whether the probability of having road deaths is equal between the 20 years. The $\alpha = 0.05$. The coding in R as below:

```
#Chi-square-test test that the road deaths is equal in the 20 years
x = main.dat$`Road Deaths`
output = chisq.test(x, correct = FALSE)
output
qchisq(0.05,19)
```

The output:

```
> output

        Chi-squared test for given probabilities

data:  x
X-squared = 540.37, df = 19, p-value < 2.2e-16

> qchisq(0.05,19)
[1] 10.11701
```

The H0 and H1 are:

H_0: $p_{(1997)} = p_{(1998)} = p_{(1999)} = \ldots . = p_{(2016)}$

H_1: at least one of the probability is different

Data extracted from output:

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

$$\boldsymbol{X}^2_{(0.05,19)} = 10.11701$$

$\boldsymbol{X}^2 = 542.37$

$p - value = 2.2 \times 10^{-16}$

From the output, the p-value < α ($2.2 \times 10^{-16} < 0.05$) and $\boldsymbol{X}^2 > \boldsymbol{X}^2_{(0.05,19)}$ (542.37 > 10.11701). I reject H0. At 95% confidence interval, there is enough evidence to conclude that the probability of having road deaths is not equal in the 20 years.

Discussion and Conclusion

Discussion

The mean number of death by road accident in Malaysia from 1997 until 2016 is 6413.95. From the number, it means every year we have about 6414 number of deaths cause by road accident, the number is round up. By using hypothesis sample of 1 sample, I test to see wether the number of road death is equal to 6000 every year. As the result suggest, the number of death is not equal to 6000, means every year the number can be below or higher than 6000.

After that, I used Pearson's product-moment correlation to check is there any relationship between the number of registered vehicle and road crashes. From the result, I obtained the value of r which will determine the correlation is equal to 0.9919396. The value is really close to 1 which means that these two variables have relatively very strong positive linear relationship between each other. Then, I can conclude that as more people buy car the more road crashes will occur.

For the next test, I used regression model to predict and explain value and impact on dependent variable from independent variable. In this study, I used the variable road crashes as independent and road deaths as dependent. This test is to check whether the two variables that I choose have realtionship or not. From the result, I could say that the number of road crashes affect the number of road death.

Lastly, I used Chi-square test with one way contingency to study is the probability of having road death caused by road accidents are equal among the 20 years. From the result, I can conclude the probability of having death caused by accidents are not equal among the 20 years. Therefore, in different years we will have different numbers of road deaths.

Conclusion

In conclusion, the objectives for the study were successfully accomplished through several test. I can conclude that even the number of registered car will affect the number of car crash, it does not mean we will have increasingly the number of car crash as the number of registered vehicle increase. We can have more people buying their desired car and still reduce the number of car crash by being a rational road user. We need to apply what we learn from our driving school such as check the car condition before use it, check the side mirror and most importantly wear the seatbelt. In addition, we must avoid reckless driving such as drive while being drunk or even drive at the emergency lane. Even you follow the procedure but being reckless while driving might also cause car accidents. Before I close this paraghraph, I want to remind that remember your loved ones before you begin your journey with your car.