



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143 – PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

SECTION : 04 - 1SECR

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

NAME : CHE NURUL FATHIAH BT MOHD HISHAMUDIN

STUDENT ID : A19EC0031

LECTURER : DR. SUHAILA BINTI MOHAMAD YUSUF

SUBMISSION DATE : 27TH JUNE 2020

Table of Contents

INTRODUCTION.....	1
HYPOTHESIS TESTING 2-SAMPLE.....	2
CORRELATION.....	3
REGRESSION.....	4
CHI SQUARE TEST OF INDEPENDENCE.....	6
CONCLUSION.....	7
REFERENCE.....	8

INTRODUCTION

For Project 2, we were assigned to analyze a secondary data that we have chose. The data set I chose is regarding the topic of Malaysia crime cases index by state. All 14 states in Malaysia crime case in 2017 were recorded along with the crime categories. This data set was collected by Ministry of Home Affairs.

Thus, I decided to look into more detail on crime cases in Johor, Kelantan, Selangor and Terengganu by analyzing some data related to them. The purpose of my study is to study the relationship between states, the type of crime cases, the total number of crime cases and the number of residents by state.

Selection of variables:

Variable	Level of measurement
States	Nominal (Johor, Kelantan, Selangor, Terengganu)
Type of crime	Nominal
Number of residents by state	Ratio
Number of property crime by state	Ratio
Number of violent crime by state	Ratio
Total number of crime by state	Ratio

Hypothesis testing 2-sample was conducted to determine if there is any difference in mean of total crime between Terengganu and Kelantan at the 0.05 significance level (95% CI). Other than that, correlation analysis was also conducted to measure the statistical relationship between number of residents and total number of crime cases in all four states. Besides that, I conduct a regression analysis to examine the linear relationship between the property crime cases and the total number of crime cases at 95% confidence interval. The last test that was being conducted to test whether a relationship exists between states and type of crime (stealing cars, stealing motorcycle) is Chi-Square test of independence with $\alpha=0.05$.

HYPOTHESIS TESTING 2-SAMPLE

(TEST ON MEAN, VARIANCE UNKNOWN, ASSUME UNEQUAL VARIANCE)

The hypothesis testing 2-sample was conducted to determine if there is any difference in mean of total crime between Terengganu and Kelantan at the 0.05 significance level (95% CI).

State	Type of crime												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Terengganu	5	64	0	79	0	76	130	365	146	914	22	0	456
Kelantan	13	114	2	141	0	78	252	547	722	1983	98	0	570

Mean sample Terengganu, $\bar{X}_1 = 173.6$

Mean sample Kelantan, $\bar{X}_2 = 347.7$

Standard deviation Terengganu, $s_1 = 264.5$

Standard deviation Kelantan, $s_2 = 549.1$

Sample size Terengganu, $n_1 = 13$

Sample size Kelantan, $n_2 = 13$

I. Hypothesis statement

$$H_0 : \mu_1 = \mu_2$$

(Null hypothesis: No difference in mean of total crime between Terengganu and Kelantan)

$$H_1 : \mu_1 \neq \mu_2$$

(Alternative hypothesis: There is difference in mean of total crime between Terengganu and Kelantan)

II. At, $\alpha = 0.05$. The test statistic is

```
> v
[1] 17.28428
> t.alpha
[1] -2.109816
> t0
[1] -1.029931
```

#v = degree of freedom

#t.alpha = critical value

#t0 = test statistic

Degree of freedom = $17.3 \approx 17$

Critical value = -2.11 (left tail), 2.11 (right tail)

Test statistic = -1.03

Therefore, using $\alpha = 0.05$, we reject the null hypothesis if $t_0 < -2.11$ or $t_0 > 2.11$

III. Conclusion

Since, $t_0 = -1.03 > t_{\alpha} = -2.11$, we fail to reject the null hypothesis

IV. Discussion

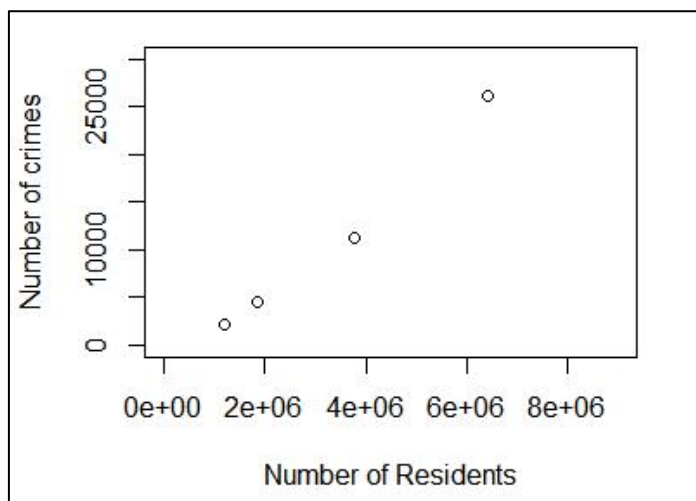
At 95% confidence interval, there is insufficient evidence to conclude that the mean of crime cases in Terengganu is different from the mean of crime cases in Kelantan.

CORRELATION

Correlation analysis was conducted to measure the statistical relationship between number of residents and total number of crime cases in Johor, Kelantan, Selangor and Terengganu.

State	Number of residents	Total number of crime cases
Johor	3 770 300	11 307
Kelantan	1 847 400	4 520
Selangor	6 401 000	26 069
Terengganu	1 221 600	2 257

I. Scatter plot



II. Sample correlation coefficient, r

```
> cor(x,y)
[1] 0.9930402
```

$$r = 0.993$$

III. Conclusion

It can be seen that the total number of crime cases increases as the number of residents increases. The scatter plot and correlation analysis of the data indicates that there is strong positive relationship between the total number of crime cases and the number of residents.

REGRESSION

Regression analysis was conducted to examine the linear relationship between the property crime cases and the total number of crime at 95% confidence interval.

State	Number of property crime cases	Total number of crimes cases
Johor	8729	11 307
Kelantan	3920	4 520
Selangor	19599	26 069
Terengganu	1903	2 257

I. Null and alternative hypothesis

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

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

II. Approximate R^2 value

```
> xx <- c(8729,3920,19599,1903)
> yy <- c(11307,4520,26069,2257)
>
> cor (xx,yy)
[1] 0.9998361
```

$$R^2 = r^2 = 0.9998361^2 = 0.99967$$

99.97% of the variation total number of crime cases is explained by variation in number of property crime cases

III. Estimated regression equation

```
> model
Call:
lm(formula = yy ~ xx)

Coefficients:
(Intercept)          xx
    -543.833         1.357
```

$$\hat{y} = -543.833 + 1.357 x$$

IV. Summary

```
> summary(model)

Call:
lm(formula = yy ~ xx)

Residuals:
    1      2      3      4 
9.305 -253.934  25.354 219.274 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -543.83329   190.14892   -2.86  0.103600
xx            1.35657     0.01737   78.09  0.000164 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 238 on 2 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9995 
F-statistic: 6099 on 1 and 2 DF,  p-value: 0.0001639
```

$$b_0 = -543.833 \quad S_E = 238 \quad S_{b_1} = 0.017 \quad \text{degree of freedom} = 2$$

p-value = 0.0001639 which is less than 0.05 indicates strong evidence against the null hypothesis.

V. Critical value, test statistic

```
> t = ((1.357-0)/0.017)
> t
[1] 79.82353
> qt(0.975,2)
[1] 4.302653
```

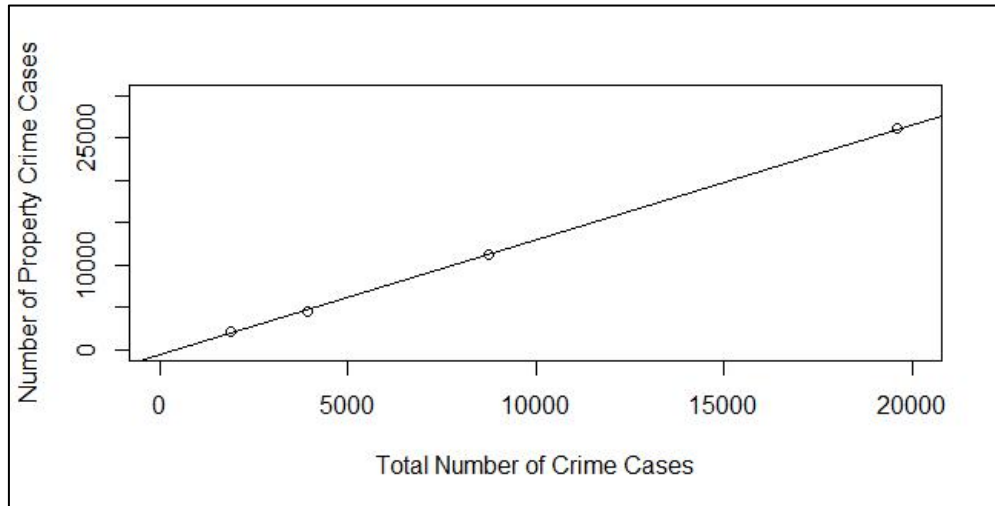
Critical value, qt = -4.303 (left tail), 4.303 (right tail)

Test statistic, t = 79.8

VI. Decision

Since $t = 79.8 > qt = 4.303$, thus we reject the null hypothesis

VII. Conclusion



At 95% confidence interval, there is sufficient evidence that linear relationship does exists between the number of property crime and the total number of crime cases. Thus, number of property crime cases do affects total number of crime cases.

CHI SQUARE TEST OF INDEPENDENCE

Chi-Square Test of Independence was conducted to test whether a relationship exists between states and type of crime (stealing cars, stealing motorcycle) using a significance level of 0.05

Type of crime	States			
	Johor	Kelantan	Selangor	Terengganu
Stealing cars	1 097	722	2 586	146
Stealing motorcycle	4 104	1 983	7 945	914

I. Null and alternative hypothesis

H_0 : No relationship exists between states and type of crime

H_1 : Exists relationship between states and type of crime

II. Critical value, test statistic, degree of freedom

```
> x2.alpha  
[1] 7.814728
```

```
> chisq.test(d, correct = FALSE)  
  
Pearson's Chi-squared test  
  
data: d  
X-squared = 94.583, df = 3, p-value < 2.2e-16
```

Degree of freedom, $df = 3$

Critical value, $x2.alpha = 7.815$

Test statistic, $X^2 = 94.58$

$p\text{-value} < 2.2e-16$ indicates that there is strong evidence against the null hypothesis.

III. Decision

Since test statistic value $>$ critical value, thus we reject the null hypothesis.

IV. Conclusion

At 95% confidence interval, there is sufficient evidence to conclude that there exists relationship between states and type of crime.

CONCLUSION

From the tests conducted to study the relationship between states, type of crime, total number of crimes and number of residents by state, a few conclusion were obtained. By using hypothesis testing on 2-sample to observe whether there are any difference in mean of total number of crime cases between Terengganu and Kelantan using 0.05 significance level, conclusively there is no significant difference between them.

Other than that, to measure the statistical relationship between number of residents and the total number of crime cases, Correlation testing was conducted and came to a conclusion that there is a strong positive relationship between the variables. It was statistically proved that the total number of crime cases increases as the number of residents increases.

In addition, Regression testing was also conducted to examine the relationship between the number of property crime and the total number of crime cases using 0.05 significance level. In conclusion, there is sufficient evidence to conclude that there is linear relationship between the variables and the number of property crime cases do affects total number of crime cases in each state.

Lastly, by conducting Chi-Square Test of Independence to test whether a relationship exists between states and type of crime (stealing cars, stealing motorcycle) at 95% confidence interval, conclusively there is sufficient evidence to conclude that there exists relationship between states and type of crime.

For the overall conclusion, we cannot analyze a set of data just by the value of the variables, even though data of the number of crime cases in one state may defer from the other.

REFERENCE

The data set. http://www.data.gov.my/data/ms_MY/dataset/statistik-jenayah-indeks-per-100-000-penduduk-mengikut-negeri-dan-kes-tahun-2017/resource/e1d2a7c9-c2a2-4720-bb1b-c3f146e96386