UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Project 2: Inferential statistics

SUBJECT NAME:	PROBABILITY AND STATISTICAL DATA ANALYSIS
SUBJECT CODE:	SECI 2143-04
SEMESTER:	2 - 2019/20
TOPIC TITLE:	CRIME STATISTIC IN MALAYSIA
STUDENT' INFO:	LIEW WEI XIAN - A19EC0070
LECTURER:	DR. SUHAILA MOHAMAD YUSUF

Table of Content

Introduction

This project uses a secondary data and the dataset I have chosen is namely Statistik Jenayah Malaysia 2019, which is collected by different government departments of Malaysia and collated by the Department of Statistics, Malaysia. The purpose of these data is to present the infographic and summary of findings of selected crime statistics. Besides, it is also to show statistical tables at Malaysia and states level regarding to selected crime statistics. Data from 2016,2017 and 2018 are featured in this dataset.

There is a variety of data that can be found in this dataset and I am interested in conducting a study on certain variables which are the number of drug addicts (male/female) from 2017 and 2018, number of crime and population of each state of Malaysia of 2018 and number of arrested cases of accepting and giving bribery of 2018. The purposes of the study are as followed:

- To investigate whether the proportion of male drug addicts in 2018 is the same with male drug addicts in 2017 in Malaysia.
- To study whether the crime index ratio is the same for all states of Malaysia in the year of 2018.
- To determine whether there is a relationship between male and female drug addicts of 2017 and 2018 in Malaysia.
- To investigate whether there appears to be a linear relationship between the number of population in each state with total crime of Malaysia in 2018.
- To determine whether the number of arrested cases of accepting bribery is dependent on the number of arrested cases of giving bribery in Malaysia, 2018.

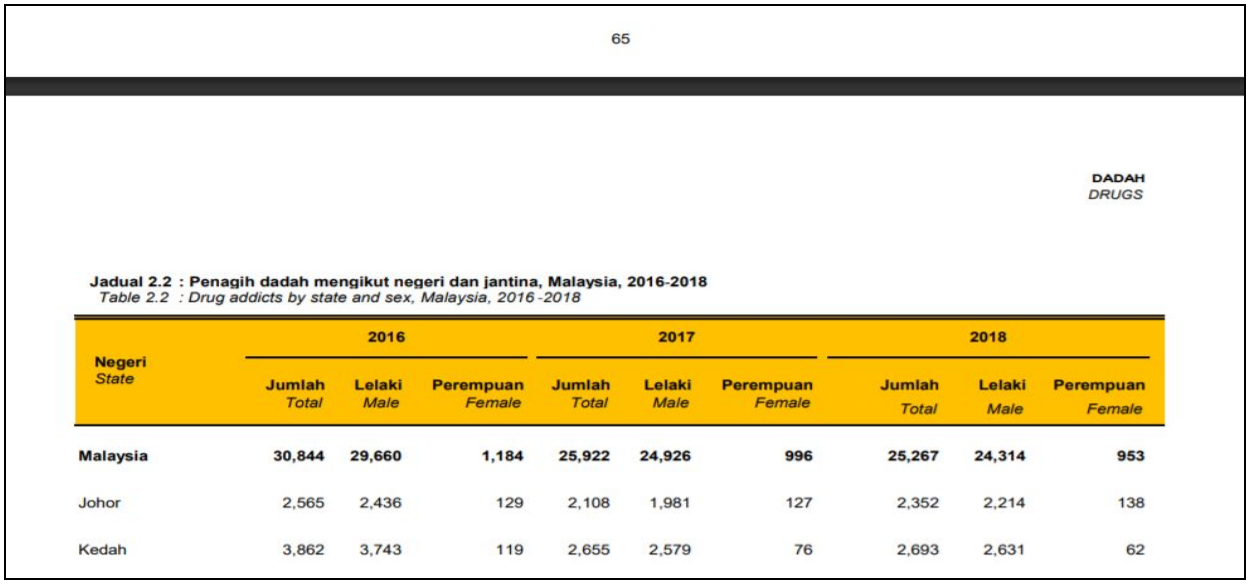
The target population is people who go against the law in Malaysia. These studies are done through hypothesis testing, chi-square testing (1 and 2 way contingency), correlation test and regression test by using R-Studio. The following discussion will be on data analysis and results (of those studies), discussion and conclusion, and last but not least the reference.

Data Analysis and Results

1. HYPOTHESIS TESTING (TWO-SAMPLE TEST)

Test:

To investigate whether the proportion of male drug addicts in 2018 is the same with male drug addicts in 2017 in Malaysia with a significance level of 0.05.



65

DADAH
DRUGS

Jadual 2.2 : Penagih dadah mengikut negeri dan jantina, Malaysia, 2016-2018
Table 2.2 : Drug addicts by state and sex, Malaysia, 2016-2018

Negeri *State*	2016			2017			2018		
	Jumlah *Total*	Lelaki *Male*	Perempuan *Female*	Jumlah *Total*	Lelaki *Male*	Perempuan *Female*	Jumlah *Total*	Lelaki *Male*	Perempuan *Female*
Malaysia	**30,844**	**29,660**	**1,184**	**25,922**	**24,926**	**996**	**25,267**	**24,314**	**953**
Johor	2,565	2,436	129	2,108	1,981	127	2,352	2,214	138
Kedah	3,862	3,743	119	2,655	2,579	76	2,693	2,631	62

Hypothesis statement:

H_0 : p1 = p2 [proportion of male drug addicts in 2018 = proportion of male drug addicts in 2017]

H_1 : p1 ≠ p2 [proportion of male drug addicts in 2018 ≠ proportion of male drug addicts in 2017]

Test statistic:

Number of male drug addicts in 2018, x1 = 24314

Number of male drug addicts in 2017, x2 = 24926

Total number of drug addicts in 2018, n1 = 25267

Total number of drug addicts in 2017, n2 = 25922

z = 0.4171615

Critical value:

$z_{0.025}$ = -1.959964

```
> #Test 1: 2-sample
> x1 = 24314
> x2 = 24926
> n1 = 25267
> n2 = 25922
> phat1 <- x1/n1
> phat2 <- x2/n2
> pbar = (x1+x2)/(n1+n2)
> qbar = 1-pbar
> z = ((phat1-phat2)-0)/sqrt((pbar*qbar/n1)+(pbar*qbar/n2))
> alpha = 0.05
> z.alpha = qnorm(alpha/2)
> z
[1] 0.4171615
> z.alpha
[1] -1.959964
>
```

Decision:

H_0 was not rejected since 0.4171615 < -1.959964.

Conclusion:

There is sufficient evidence at 0.05 significance level to support that the proportion of male drug addicts in 2018 is the same as male drug addicts in 2017 in Malaysia.

2. CHI-SQUARE GOODNESS OF FIT TEST (ONE -WAY CONTINGENCY)

Test:

To study whether the crime index ratio is the same for all states of Malaysia in the year of 2018 at 5% significance level.

Hypothesis statement:

H_0 : c1 = c2 = c3 = c4 = c5 = c6 = c7 = c8 = c9 = c10 = c11 = c12 = c13 = c14

[The crime index ratio is the same for all states of Malaysia in the year of 2018.]

H_1 : At least one state's crime index ratio is different from the other state's crime index ratio in the year of 2018, Malaysia.

State in Malaysia	Crime Index Ratio in 2018
Johor	275.7
Kedah	287.6
Kelantan	214.3
Melaka	303.6
Negeri Sembilan	327.1
Pahang	215.3
Perak	204.8
Perlis	222.1
Pulau Pinang	284.6
Sabah	153.9
Sarawak	208.8
Selangor	330.8
Terengganu	148.4
W.P. Kuala Lumpur	642.6

Test statistic:

$\chi^2 = 703.7804$

Critical value:

Degree of freedom = 14-1 = 13

$\chi^2_{13,0.05} = 22.36203$

```
> #Test 2: Chi-square goodness of fit test
> noAcc <- c(275.7, 287.6, 214.3, 303.6, 327.1, 215.3, 204.8,
+            222.1, 284.6, 153.9, 208.8, 330.8, 148.4, 642.6)
> expprob <- sum(noAcc)/14
> expAcc <- c(expprob, expprob, expprob, expprob, expprob, expprob, expprob, expprob,
+             expprob, expprob, expprob, expprob, expprob, expprob)
> exp <- ((noAcc-expAcc)^2)/expAcc
> x2 <- sum(exp)
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df = 13, lower.tail=FALSE)
> x2.value <- pchisq(x2, df= 13, lower.tail = FALSE)
> print(x2)
[1] 703.7804
> print(x2.alpha)
[1] 22.36203
```

Decision:

H_0 was rejected since 703.7804 > 22.36203

Conclusion:

There is sufficient evidence at 0.05 significance level to support that at least one state's crime index ratio is different from the other state's crime index ratio in the year of 2018, Malaysia.

3. CHI - SQUARE INDEPENDENCE TEST (TWO-WAY CONTINGENCY)

Test:

To determine whether there is a relationship between male and female drug addicts of 2017 and 2018 in Malaysia at 5% significance level.

Hypothesis statement:

H_0: There is no relationship between male and female drug addicts of 2017 and 2018 in Malaysia.

H_1: There is a relationship between male and female drug addicts of 2017 and 2018 in Malaysia.

Gender\Year	2017	2018
Male	24,926	24,314
Female	996	953

Test statistic:

$\chi^2 = 0.17402$

Critical value:

Degree of freedom = (2-1)(2-1) = 1

$\chi^2_{1,0.05} = 3.841459$

```
> #///////Test 3: CHi-aquare test of independence
> male <- c(24926,24314)
> female <- c(996, 953)
> d <- data.frame(male, female)
> chisq.test(d, correct = FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 0.17402, df = 1, p-value = 0.6766

> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail = FALSE)
> print(x2.alpha)
[1] 3.841459
```

Decision:

Failed to reject H_0 since 0.17402 < 3.841459.

Conclusion:

There is sufficient evidence at 0.05 significance level to support that there is no relationship between male and female drug addicts of 2017 and 2018 in Malaysia.

4. CORRELATION TEST

Test:

To investigate whether there appears to be a linear relationship between the number of population in each state with total crime of Malaysia in 2018 at 5% significance level.

Negeri *State*	Tahun *Year*	Nisbah jenayah Indeks *Crime index ratio*	Jenayah indeks *Crime Index*	Penduduk *Population* ('000)
Malaysia	**2016**	**355.2**	**112,355**	**31,633.4**
	2017	**309.7**	**99,162**	**32,022.5**
	2018	**273.8**	**88,662**	**32,382.3**
Johor	2016	354.4	12,941	3,651.8
	2017	305.8	11,307	3,697.0
	2018	275.7	10,338	3,749.4
Kedah	2016	351.0	7,440	2,119.7
	2017	315.3	6,759	2,143.9
	2018	287.6	6,221	2,163.0
Kelantan	2016	253.1	4,548	1,796.7
	2017	247.1	4,520	1,829.0
	2018	214.3	3,987	1,860.5

Hypothesis statement:

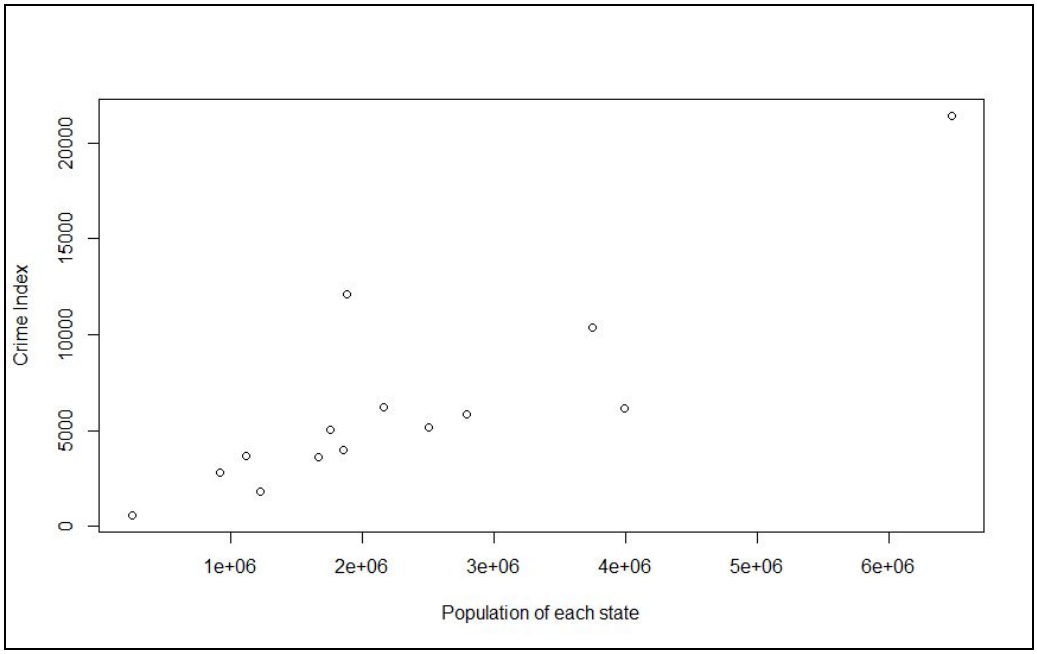
$H_0: \rho = 0$

[There is no linear correlation between the number of population in each state with total crime of Malaysia in 2018]

$H_1: \rho \neq 0$

[There is a linear correlation between the number of population in each state with total crime of Malaysia in 2018]

Plot:



Test Statistic:

Pearson's correlation coefficient, r = 0.8598453

t = 5.834

Critical Value:

Degree of freedom = 14-2 = 12

$t_{12,0.05}$ = (-2.179, 2.179)

```
> #//////Test 4: CORRELATION TEST///////
> x <- c(3749400, 2163000, 1860500, 922400, 1122900, 1664700, 2503500, 253500, 1762800, 3997400
7200)
> y <- c(10338, 6221, 3987, 2800, 3673, 3584, 5128, 563, 5017, 6151, 5830, 21420, 1823, 12127)
> plot(x,y, xlab = "Population of each state", ylab = "Crime Index")
> res <- cor.test(x, y, method = "pearson")
> res

        Pearson's product-moment correlation

data:  x and y
t = 5.834, df = 12, p-value = 8.035e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6055087 0.9548202
sample estimates:
      cor
0.8598453
```

Decision:

Reject H_0 since 5.834 > 2.179.

Conclusion:

There is sufficient evidence at 95% confidence level to support that there is a linear correlation between the number of population in each state with total crime of Malaysia in 2018. And since the Pearson's correlation coefficient, r is 0.8598453 > 0.8, it is a strong positive linear correlation. The more the population of one state, the higher the crime index for that particular state.

5. REGRESSION TEST

Test:

To determine whether the number of arrested cases of accepting bribery cases is dependent on the number of arrested cases of giving bribery cases in Malaysia, 2018 at 5% significance level.

Dependent variables (y): Number of arrested cases of accepting bribery cases in Malaysia, 2018

Independent variables (x): Number of arrested cases of giving bribery cases in Malaysia, 2018

Hypothesis Statement:

$H_0 : \beta_1 = 0$

[Number of arrested cases of accepting bribery cases is not dependent on number of giving bribery cases in Malaysia, 2018]

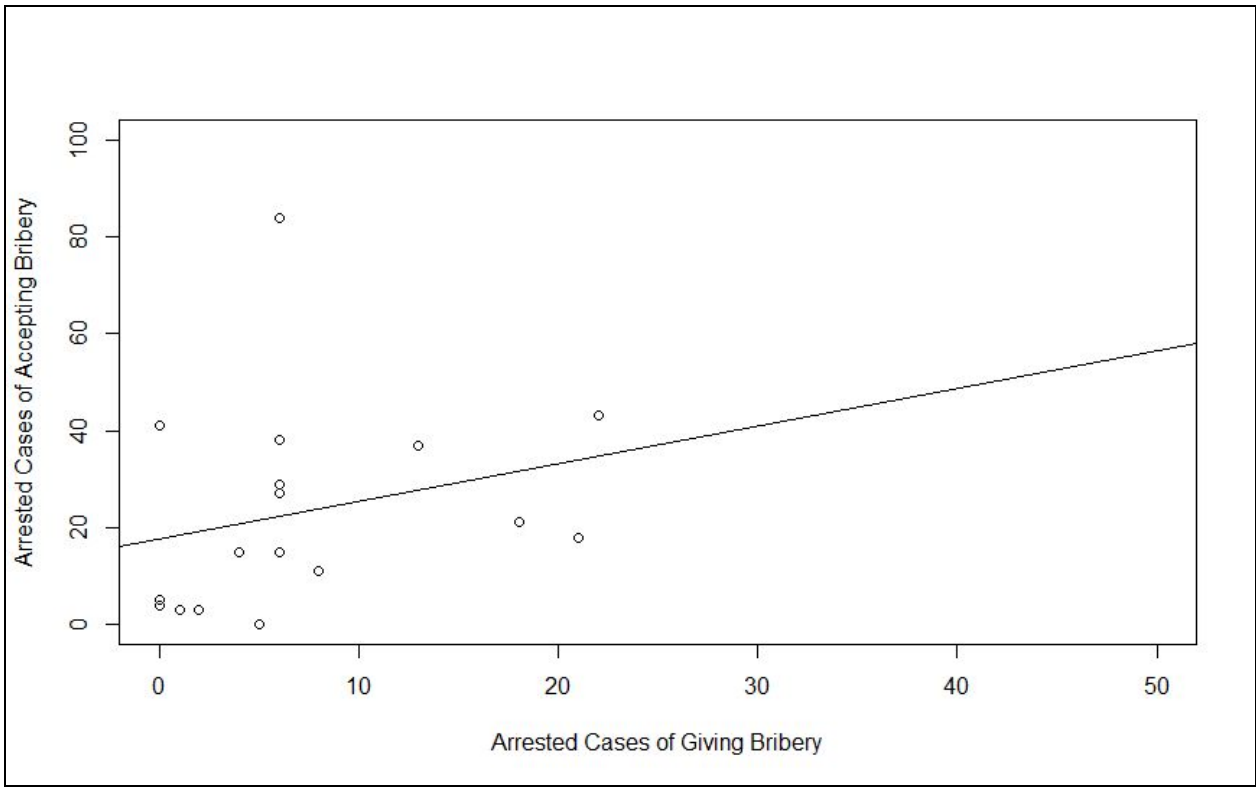
$H_1 : \beta_1 \neq 0$

[Number of arrested cases of accepting bribery cases is dependent on number of giving bribery cases in Malaysia, 2018]

Location	Giving Bribery(x)	Accepting Bribery(y)
Johor	22	43
Kedah	18	21
Kelantan	5	0
Melaka	0	5
Negeri Sembilan	6	29
Pahang	2	3
Perak	6	15
Perlis	1	3
Pulau Penang	0	41
Sabah	13	37

Sarawak	21	18
Selangor	6	38
Terengganu	6	27
W.P. Kuala Lumpur	8	11
W.P. Labuan	0	4
W.P. Putrajaya	4	15
Ibu Pejabat	6	84

Plot:



Test Statistic:

Pearson's correlation coefficient, r = 0.2588941

$\hat{y} = 17.511 + 0.7767x$

p-value = 0.3157

```
Call:
lm(formula = dp ~ idp)

Residuals:
    Min      1Q  Median      3Q     Max
-21.395 -13.511  -7.171   8.401  61.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.5110     7.4998   2.335   0.0339 *
idp           0.7767     0.7482   1.038   0.3157
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.21 on 15 degrees of freedom
Multiple R-squared:  0.06703,   Adjusted R-squared:  0.004828
F-statistic: 1.078 on 1 and 15 DF,  p-value: 0.3157
```

Decision:

Failed to reject H_0 since 0.3157 > 0.05

Conclusion:

There is sufficient evidence at 95% confidence level to support that the number of arrested cases of accepting bribery is not dependent on the number of arrested cases of giving bribery in Malaysia, 2018. The Pearson's correlation coefficient, r = 0.2588941 also shows that there is a fairly weak linear relationship between those two variables.

Discussion and Conclusion

First of all, according to the first testing, 2-sample hypothesis testing to investigate whether the proportion of male drug addicts in 2018 is the same with male drug addicts in 2017 in Malaysia with a significance level of 0.05. I found that there is no different in the proportion of male drug addicts between 2017 and 2018. The result is kinda sad as the proportion is still the same between those 2 years but the population of male drug addicts has decreased which is nice to see.

Besides, from the chi-square testing I found that the crime index ratio is not the same for all states of Malaysia in the year of 2018. Crime index ratio in cities tends to be higher than the rest. Meanwhile, the study also shows that there is no relationship between male and female drug addicts of 2017 and 2018 in Malaysia.

Last but not least, during the correlation test I found that there is a strong positive linear correlation between the number of population in each state with total crime of Malaysia in 2018. The more the population of one state, the higher the crime index for that particular state. However, during the regression test I found that the number of arrested cases of accepting bribery is not dependent on the number of giving bribery in Malaysia, 2018. The result is to my surprise, as I thought it was the opposite. My assumption is that since this is done based on respective locations in Malaysia (more on states), there is a possibility that there is inter-location bribery business going on which makes the result a little bit off.

Reference

1. Statistik Jenayah Malaysia 2019 (pdf). Retrieved from https://newss.statistics.gov.my/newss-portalx/ep/epFreeDownloadContentSearch.seam?cid=34629
2. Correlation Test Between Two Variables in R. Retrieved from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
3. Lecturer's note: Chapter 5
4. Lecturer's note: Chapter 6
5. Lecturer's note: Chapter 7