UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY AND STATISTICAL DATA ANALYSIS

(SECI2143–02)

PROJECT 2

Lecturer:

DR. NOR AZIZAH ALI

Group Members:

Name	Matric No
MUHAMMAD IQMAL BIN SIS	A21EC0080
MUHAMMAD HASAN BIN CHE ABDULLAH	A21EC0077
WAN AMIRUL HAFIQ BIN WAN HUZAINI	A21EC0141
MUHAMMAD IZZUDDIN BIN SHABRIN	A21EC0083
MUHAMMAD ASHRAAF BIN SALEH	A21EC0068

Table of Content

1.0 Introduction

Throughout the second semester of our bachelor's degree studies, we have enrolled in the SECI2143 course i.e Probability And Statistical Data Analysis. As stated in the course learning outcome, we are tasked with conducting an inference statistical analysis on a chosen dataset which in our case is "Car Related Road Accidents In Malaysia (2010-2021) and Registered Vehicles In Malaysia (2010-2022)". Group 3 decided to tackle this topic since it is one that is of great concern to us as road users. Not only that, as Malaysians we are well aware of our society's dependance on road vehicles in our everyday lives, it is undeniable that we are a car-centric society first and foremost. What that entails is wide reaching and oftentimes adverse to society with environmental pollution, high cost of living, inefficient city planning and perhaps worst of all, loss of lives. We've also seen for ourselves the effects of more than 17 million vehicles on our congested roads especially in developed urban areas, highways and during peak hours, it greatly reduces quality of life and poses a safety risk to all road users, even ones not in vehicles. Thus, our aim is to analyse road safety that has always been a dire issue in Malaysia and by that we mean to identify the relationship between the amount of cars and the frequency of road accidents as well as compare the frequency of road accidents in a number of states across multiple years. With our analysis, we hope to gain a better understanding on at least a few viewpoints and determine the extent of vehicle ownership's effects on road safety. We made sure to utilise all the tools and inferential methods that we have learned in our syllabus namely Hypothesis 2 sample test, correlation and regression tests, goodness of fit test and Chi square test of independence in RStudio alongside data visualisations such as scatter plots to paint a clearer picture of the information that can be gained from our dataset.

2.0 Dataset

The dataset that we obtained from data.gov.my is a secondary data which contains total of accidents from each state in Malaysia from 2010 to 2021, car registered. The dataset is used to to study the relationship between the variables, predict the number of domestic total car-related accidents in the same years, and understand the factors of car accidents in certain states.

The dataset contains 12 variables but we only use 5 of them which is year, states, total fatal accidents, total accidents, and total registered cars. We will carry out 5 statistical tests which are Hypothesis testing 1-sample, Correlation test using variable, Regression test, Chi-Square test of independence,and Goodness of fit test.

3.0 Data Analysis

3.1 Hypothesis One-Sample Test

Hypothesis null, H_0 : total number of accidents in Malaysia in 2021 is 10% of total cars registered in Malaysia in the same year.

Alternative hypothesis, H_1: total number of accidents in Malaysia in 2021 is 10% greater than the total cars registered in Malaysia in the same year.

Tahun	Total of car accident	Total of car registered
2021	182616	508911

H_0 : p = 0.10

H_1 : p > 0.10

α = 0.1 (right-tail)

$$\overline{p} = \frac{182616}{508911}$$
$$= 0.3588$$

$$\sigma = \sqrt{\frac{(182616 - 0.3588)^2}{508911}}$$
$$= 255.986$$

$$z = \frac{\overline{p} - \mu}{\sigma/\sqrt{n}}$$
$$= \frac{0.3588 - 0.10}{255.986/\sqrt{508911}}$$
$$= 0.721$$

where:

$\bar{p}$ = sample mean

μ = hypothesized population mean in the null hypothesis

σ= standard deviation

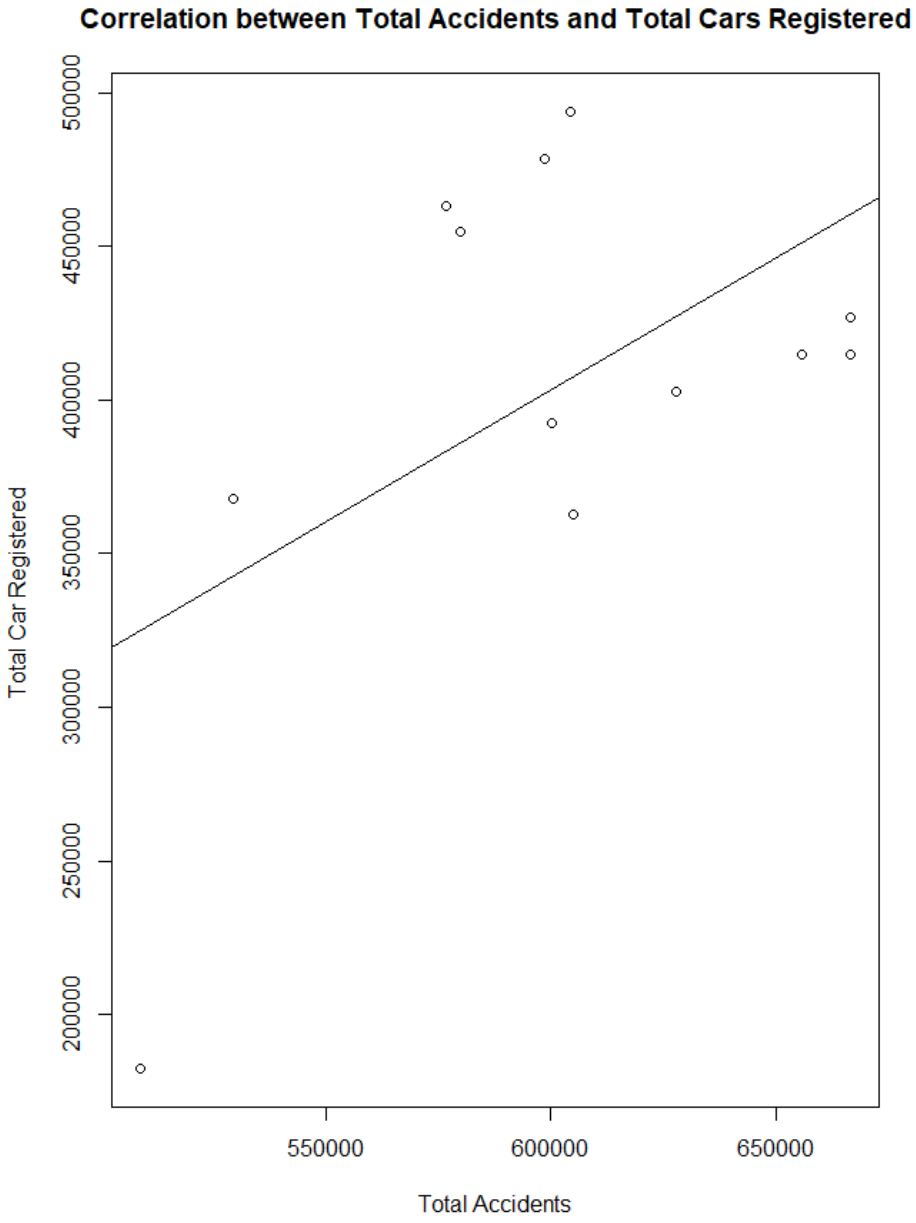
n = sample size

P(z < 0.721) = 0.4994

:- Since 0.4994 > 0.10, therefore fail to reject H_0. The mistake of failing to reject the null hypothesis when it is false is called type II error.Based on that this hypothesis testing is a type II error. Fail to reject p ≤ 0.10 and also fail to support p > 0.10 when in reality p > 0.10.

3.2 Correlation Test

In the correlation test, we will conduct a test to investigate whether there is a linear relationship between the total number of accidents with the total number of cars registered from 2010-2021 with a sample of xx. We conduct the test with a significance level of 0.05



In this test, we will use the Pearson's product-moment correlation coefficient to calculate the r, which is a value used to measure the strength of the linear relationship in the sample observed. The calculation to find r is as below:

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

$$= \frac{2943903615606 - \frac{35048716940652}{12}}{\sqrt{[(2036235846872) - (4854362)^2/12][(4371051834480) - (7220046)^2/12]}}$$

$$= 0.524$$

where:

r = sample correlation coefficient

n = sample size

x = value of independent variable

y = value of dependant variable

Tahun	Total_Car(y)	Total_Accidents(x)	x*y	x^2	Y^2
2010	605156	362938	219634108328.00	131723991844.00	366213784336.00
2011	600123	392256	235401847488.00	153864769536.00	360147615129.00
2012	627733	402563	252702079679.00	162056968969.00	394048719289.00
2013	655744	414690	271930479360.00	171967796100.00	430000193536.00
2014	666487	414675	276375496725.00	171955355625.00	444204921169.00
2015	666667	426957	284638142319.00	182292279849.00	444444888889.00
2016	580085	454616	263715922360.00	206675707456.00	336498607225.00
2017	576625	462952	266949697000.00	214324554304.00	332496390625.00
2018	598714	478332	286384065048.00	228801502224.00	358458453796.00
2019	604287	493745	298363684815.00	243784125025.00	365162778369.00
2020	529514	368022	194872801308.00	135440192484.00	280385076196.00
2021	508911	182616	92935291176.00	33348603456.00	258990405921.00
TOTAL	7220046	4854362	2943903615606.00	2036235846872.00	4371051834480.00
	POWER	sum x*sum y			
	2.00	35048716940652.00			

Figure ….: Calculation of the variables for Pearson's product-moment correlation coefficient.

From the Pearson's product-moment correlation coefficient, we get the value of r = 0.524, which indicates that the relationship between total accidents and total cars registered is a moderate positive linear relationship. After obtaining the value of r, we will continue the testing by following the below steps

i) State the hypothesis statement

$h_0: \ \rho = 0$ (no linear correlation)

$h_1: \ \rho \neq 0$ (linear correlation exist)

ii) Find the critical value

$\alpha = 0.05, \ d.f = n - 2 = 12 - 2 = 10$

$t_{\alpha/2 = 0.025,10} = \pm 2.228$

iii) Calculate the test statistic

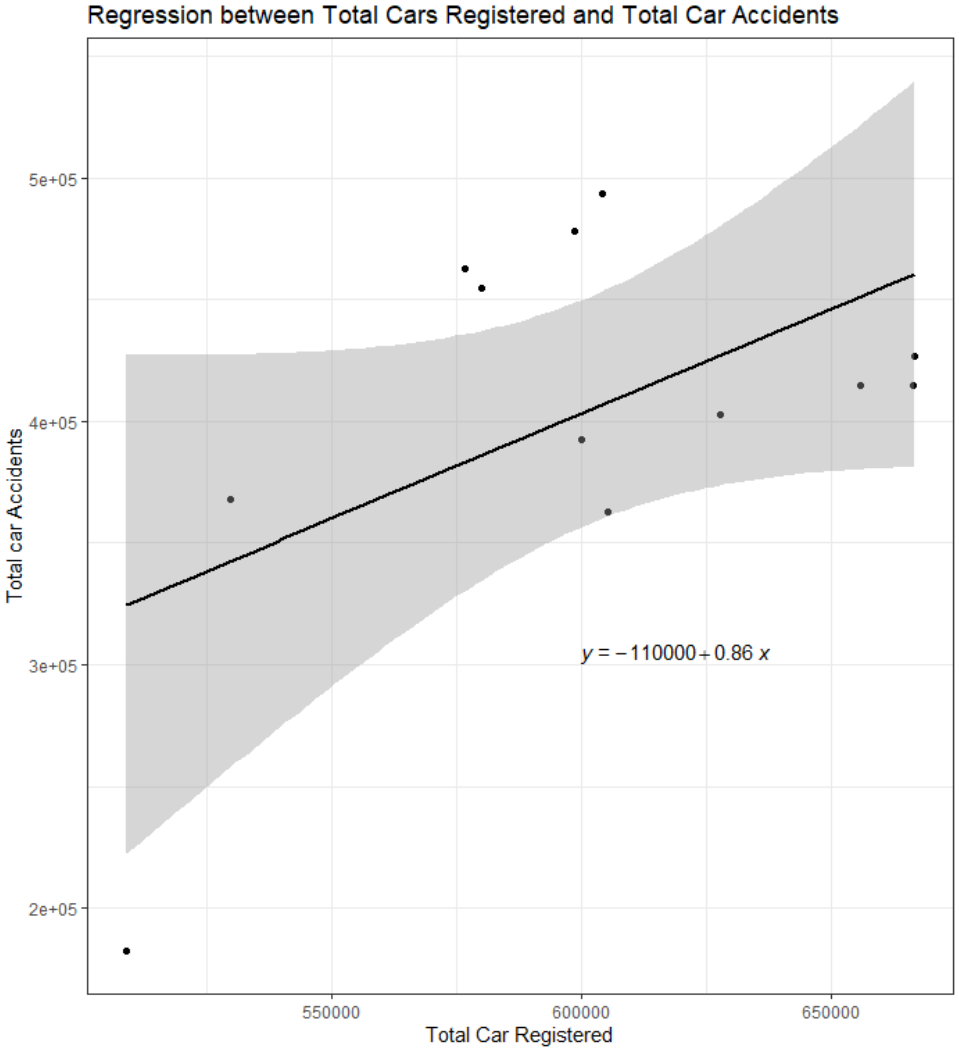
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.524}{\sqrt{\frac{1-0.524^2}{12-2}}} = 1.95$$

iv) Conclusion

Since $t = 1.95 < t_{0.025,10} = 2.228$, fail to reject h_0. There is no sufficient evidence to conclude that there is a linear relationship between total cars registered and total accidents at the 5% level of significance.

3.3 Regression Test

The simple regression test we conducted is to determine if the total number of newly registered cars in Malaysia from the years 2010-2021 could be used to predict the number of domestic total car-related accidents in the same years. As such, we set the total number of newly registered cars as the independent variable or 'x' while the number of domestic total car-related accidents is set as the dependent variable, 'y'. As usual, we set the confidence level as $\alpha = 0.05$. This test is partly used to detect if a linear relationship is present between the two aforementioned variables.



Figure…: Regression Table

The accompanying regression equation in the figure above, $y = -110000 + 0.86x$ is that of the regression line which is represented by:

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

where:

$\hat{y}$ = estimated/predicted y value

b_0 = estimate of regression intercept

b_1 = estimate of regression slope

x = value of independent variable

We obtained the value for both b_0, b_1 and R^2 through summarising the dataset in RStudio as well as confirming it by using our own calculations in Microsoft Excel, which we were able to derive the same approximate values. As shown in the scatter plot, the data points are somewhat in linear relation but with a wider range for errors. The value for intersection coefficient, b_0 is -110000 which means the hypothetical intersection point (value) of total car accidents in Malaysia when the number of registered cars in the country reaches 0 is in the negative mark. This is obviously impossible but it's indicative of the relationship between the two. The slope coefficient, b_1 at a value of 0.86 indicates that a single unit degree change in the independent variable would lead to a 0.86 change in the dependent variable in either direction. As seen in the figures below, the square of R value which is the coefficient of determination, R^2 has a value of 0.2748 meaning that there exist only a weak linear relation between the two variables where some but not all of the variation in *y* is explained by variation in *x*.

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

Residuals:
    Min      1Q  Median      3Q     Max
-142179  -38388  -17657   70566   86966

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.127e+05  2.665e+05  -0.423   0.6814
x            8.596e-01  4.416e-01   1.947   0.0802 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72510 on 10 degrees of freedom
Multiple R-squared:  0.2748,    Adjusted R-squared:  0.2023
F-statistic: 3.789 on 1 and 10 DF,  p-value: 0.0802
```

Figure … : Summary of regression in RStudio

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$sb_1 = 0.4416$

p-val = 0.0802

Coefficient of determination, $R^2 = 0.2748$

$b_0 = -110000$, $b_1 = 0.86$

t value = 1.947

TAHUN	KEMALANGAN (Y)	KENDERAAN (X)
2010	362,938	605,156
2011	392,256	600,123
2012	402,563	627,733
2013	414,690	655,744
2014	414,675	666,487
2015	426,957	666,667
2016	454,616	580,085
2017	462,952	576,625
2018	478,332	598,714
2019	493,745	604,287
2020	368,022	529,514
2021	182,616	508,911
TOTAL	4,854,362	7,220,046

Table… : Table of Total Car Accidents and Total Registered Cars in Malaysia (2010-2021)

XY	X^2	Y HAT	Y AVE	sst	sse	ssr
219,634,108,328	366,213,784,336	407,526.261	404530.17	1,729,908,328.03	1,988,113,053.06	8,976,583.54
235,401,847,488	360,147,615,129	403,199.954		150,655,167.36	119,770,127.13	1,769,465.98
252,702,079,679	394,048,719,289	426,933.184		3,869,744.69	593,905,890.57	501,895,206.21
271,930,479,360	430,000,193,536	451,011.110		103,222,213.36	1,319,223,023.20	2,160,478,082.37
276,375,496,725	444,204,921,169	460,245.666		102,917,643.36	2,076,685,604.48	3,104,216,871.82
284,638,142,319	444,444,888,889	460,400.392		502,962,853.36	1,118,460,463.84	3,121,482,071.07
263,715,922,360	336,498,607,225	385,975.525		2,508,590,700.69	4,711,514,766.73	344,274,716.16
266,949,697,000	332,496,390,625	383,001.350		3,413,110,610.03	6,392,106,421.60	463,489,943.34
286,384,065,048	358,458,453,796	401,988.794		5,446,710,603.36	5,828,285,084.66	6,458,574.44
298,363,684,815	365,162,778,369	406,779.279		7,959,286,486.69	7,563,036,589.90	5,058,507.30
194,872,801,308	280,385,076,196	342,505.290		1,332,846,233.36	651,102,499.44	3,847,085,350.35
92,935,291,176	258,990,405,921	324,795.194		49,245,897,367.36	20,214,923,165.07	6,357,665,889.38
2,943,903,615,606	4,371,051,834,480					

Table… : Calculations to find Explained & Unexplained Variations and R^2

done in Microsoft Excel

B1	0.859588213		
B0	-112658.703		
SST	SSE	SSR	R^2
72,499,977,951.67	52,577,126,689.68	19,922,851,261.98	0.274798032

Table… : Calculations to find R^2 done in Microsoft Excel

Test Statistic:

$$t = \frac{b_1 - \beta_1}{sb_1}$$

$$t = \frac{0.86 - 0}{0.4416}$$

$$t = 1.947$$

Critical Value:

$$df = 12 - 2 = 0$$

$$\alpha = 0.05$$

$$t_{0.05/2} = 2.228$$

Decision:

Since $t = 1.947$ of *Test Statistics* is less than $t_{0.05/2} = 2.228$ of *Critical Value*, thus we fail to reject H_0.

Conclusion:

In conclusion, by using 0.05 significance level, there is insufficient evidence to support that the total number of registered cars in Malaysia (2010-2021) has any effects on the total number of car accidents in Malaysia (2010-2021).

3.4 Goodness of Fit Test

In the goodness of fit test, we will conduct a test to investigate whether the claim that the accident occurs in Selangor has an equal proportion in the 5 years (2017-2021). We use 4190 samples that were involved in a fatal accident within this 5 years (2017-2021). We conduct the test with a significance level of 0.05.

Based on the claim that the proportion of accidents in Selangor is equal, therefore

$$\rho 2017 = \rho 2018 = \rho 2019 = \rho 2020 = \rho 2021$$

ρ = Probability for the category

Hypothesis Statement

H0 **:** $\rho 2017 = \rho 2018 = \rho 2019 = \rho 2020 = \rho 2021$

H1 **:** At least one of the five proportions is different from others.

Tahun	2017	2018	2018	2020	2021	Total
Observed accidents, O	1,047	1,004	1,008	764	367	4,190

Figure xx: Depicts Data of total death due to accident in Selangor and year.

Calculate The Expected Frequency

Calculate the expected frequency using the formula given below

$$E = n/k$$

n = Represents the total number of trials

k = Represents the number of different categories or outcomes

Tahun	2017	2018	2018	2020	2021	Total
Observed accidents, O	1,047	1,004	1,008	764	367	4,190
Expected accidents, E	838	838	838	838	838	4190

Calculate The Difference Between O and E:

O = Represents the observed frequency of an outcome

E = Represents the expected frequency of an outcome

Using this formula $\frac{(O-E)^2}{E}$

Tahun	2017	2018	2018	2020	2021	Total
Observed accidents, O	1,047	1,004	1,008	764	367	4,190
Expected accidents, E	838	838	838	838	838	4190
(O -E)2/E	52.125	32.883	34.487	6.535	264.727	390.757

Calculate The Test Statistic

$$\chi 2 = \Sigma\frac{(O-E)^2}{E} = 52.125+32.883+6.535+264.727 = 390.757$$

Using the significance level of 0.05%, we calculate the critical value
Degree of Freedom = 5 - 1 = 4

$\chi 2_{4,0.05} = 9.48$

State The Decision

Test Statistic, $\chi 2$ = 390.757

Critical Value, $\chi 2$ 4,0.05 = 9.48

Decision:

Since Test Statistic is greater than Critical Value, thus we reject the null hypothesis, ***H0*** at α = 0.05

Conclusion

There is no sufficient evidence, so we reject the claim that the fatal accidents occur in Selangor are an equal proportion on the years.

3.5 Chi-Square Test of Independence

a) Hypothesis Statement

H0 **:** Total accident of State is independent of the year of total accident
H1 **:** Total accident of State is dependent of the year of total accident

State	2017	2018	2019	2020	2021	Total
SELANGOR	154,958	163,078	168,222	123,230	60,370	669,858
NEGERI SEMBILAN	24,941	25,123	25,838	19,905	9,611	105,418
MELAKA	18,771	19,120	19,593	14,543	7,130	79,157
JOHOR	76,121	78,812	82,502	58,207	28,157	323,799
PAHANG	20,813	20,641	21,196	17,000	8,071	87,721
Total	295,604	306,774	317,351	232,885	113,339	1,265,953

Figure xx: Shows Data of State of total accident and year of total accident

b) Critical Value

Test Statistic:

$$X^2 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

o- Observed Count
e- Expected Count

Degree of Freedom, df = (r-1) (c-1)

Critical Value, α = 0.05
Degree of Freedom, df = (5-1)(5-1) =16

c) Calculate Expected Counts

Expected Count calculation formula:

$$e_i = \frac{(i^{\text{th}}\ Row\ total)\ (j^{\text{th}}\ Column\ total)}{\text{Total sample size}}$$

	2017	2017	2018	2018
State	**Observed count**	**Expected count**	**Observed count**	**Expected count**
SELANGOR	154,958	156413.95	163,078	162,324.37
NEGERI SEMBILAN	24,941	24615.43	25,123	25,545.58
MELAKA	18,771	18483.41	19,120	19,181.84
JOHOR	76,121	75608.08	78,812	78,465.09
PAHANG	20,813	20483.13	20,641	21,257.13
Total	295,604		306,774	

Figure xx: Shows Observed Count and Expected Count for the selected data (Part I)

2019	2019	2020	2020	2021	2021	
Observed count	**Expected count**	**Observed count**	Expected Count	**Observed count**	Expected Count	**Total**
168,222	167,921.01	123,230	123,227.23	60,370	59,971.45	669,858
25,838	26,426.34	19,905	19,392.72	9,611	9,437.93	105,418
19,593	19,843.20	14,543	14,561.74	7,130	7,086.82	79,157
82,502	81,170.42	58,207	59,566.14	28,157	28,989.27	323,799
21,196	21,990.03	17,000	16,137.17	8,071	7,853.54	87,721
317,351		232,885		113,339		1,265,953

Figure xx: Shows Observed Count and Expected Count for the selected data (Part II)

d) Calculate Test Statistic Value

State	2017	2017	2017
	Observed Count, o_{ij}	**Expected count, e_{ij}**	**$[(o_{ij} - e_{ij})]^2/e_{ij}$**
SELANGOR	154,958	156413.95	13.55
NEGERI SEMBILAN	24,941	24615.43	4.31
MELAKA	18,771	18483.41	4.47
JOHOR	76,121	75608.08	3.48
PAHANG	20,813	20483.13	5.31

2018	2018	2018
Observed Count, o_{ij}	**Expected count, e_{ij}**	**$[(o_{ij} - e_{ij})]^2/e_{ij}$**
163,078	162,324.37	3.50
25,123	25,545.58	6.99
19,120	19,181.84	0.20
78,812	78,465.09	1.53
20,641	21,257.13	17.86

2019	2019	2019
Observed Count, o_{ij}	**Expected count, e_{ij}**	**$[(o_{ij} - e_{ij})]^2/e_{ij}$**
168,222	167,921.01	0.54
25,838	26,426.34	13.10
19,593	19,843.20	3.15
82,502	81,170.42	21.84
21,196	21,990.03	28.67

2020	2020	2020
Observed Count, o_{ij}	**Expected count, e_{ij}**	**$[(o_{ij} - e_{ij})]^2/e_{ij}$**
123,230	123,227.23	6.23×10^{-5}
19,905	19,392.72	13.53
14,543	14,561.74	0.02
58,207	59,566.14	31.01
17,000	16,137.17	46.13

2021	2021	2021
Observed Count, o_{ij}	**Expected count, e_{ij}**	**$[(o_{ij} - e_{ij})]^2/e_{ij}$**
60,370	59,971.45	2.65
9,611	9,437.93	3.17
7,130	7,086.82	0.26
28,157	28,989.27	23.89
8,071	7,853.54	6.02

X^2	255.18

e) State the decision

Test Statistic, $X^2 = 255.18$

Critical Value, $X^2_{=16,\ \alpha\ =\ 0.05} = 26.296$

Decision:

Since Test Statistic is greater than Critical Value, thus reject the null hypothesis, *H0* at $\alpha = 0.05$

Conclusion

There is evidence of a relationship between State of the total accident and year of total accident

4.0 Conclusion

In a nutshell, we have learned various technique to satisfy the objectives of our data analysis from our chosen topics which are “Car Related Road Accidents in Malaysia (2010-2021)” and “Registered Vehicles in Malaysia (2010-2022)”. We got this data from Data.gov.my, the official website to obtain trusted data. Throughout the entire analysis, we are aware that there are various factors that contribute to car accidents in Malaysia.

For data pre-proccessing, we have done 2 methods which are data cleaning and data reduction. In which we remove 1 year from “Car Related Road Accidents in Malaysia (2010-2021)” to make it easier to calculate. We also clean the data by removing unrelated data from “Car Related Road Accidents in Malaysia (2010-2021)”.

Based on our analysis, we can conclude that the total number of accidents in Malaysia is 10% greater than the total number of cars registered in Malaysia (2010-2021). Next, there is no linear correlation between the total number of accidents in Malaysia (2010-2021) and the total number of cars registered in Malaysia (2010-2021). We also conclude that there is insufficient evidence to support that the total number of cars registered in Malaysia (2010-2021) has any effect on car accidents in Malaysia (2010-2021). We also do an analysis to see whether a fatal accident has an equal proportion, using a specific state which is Selangor and we can conclude that fatal accidents in Selangor are not in equal proportion. Lastly, we conclude that there is a relationship between car accidents and between state and year of accidents.

We would like to thank Dr. Nor Azizah Ali, our Probability and Statistical Data Analysis Lecturer for giving us the opportunity to learn various techniques to dissect the data using R programming, how to clean the data, and how to read the data, and how to perform analysis for the data. We also gained knowledge on how each variable will affect the data based on the analysis that we do.

5.0 Appendix

Malaysia Number of Registered Vehicles, 1986 – 2022 | CEIC Data
Cities 101: The hidden cost of car-centric cities (Part IV) – | B L O G (wordpress.com)