UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143 – PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

DATA ANALYSIS BASED ON INFERENTIAL STATISTICS

SECTION : 06 - 1SECJH

LECTURER'S NAME : DR. IZYAN IZZATI BINTI KAMSANI

DATE OF SUBMISSION : 30th JUNE 2021

No.	Name	Metric No.
1	AHMAD NAZRAN BIN YUSRI	A20EC0179
2	MUHAMMAD HAFIZZUL BIN ABDUL MANAP	A20EC0211
3	MEGAT IRFAN ZACKRY BIN ISMAIL	A20EC0205

Table of Contents

No.	Content	Page
1	Background ● Purpose of Study ● Interest ● Our Expectation	3
2	Dataset	4
3	Data Analysis ● Hypothesis 1-Sample Test ● Correlation Test ● Regression Test ● Chi Square Test	5
4	Conclusion	12
5	Appendix	13

1.0 BACKGROUND

A forest can be understood as an ecosystem or large area of land which is dominated and covered with trees and other woody vegetation. There is no doubt that the forest is the precious treasure and asset that is important for our continuation and survival. Forest provides a lot of functions and benefits not only to humankind, but also for other life forms. It is a home for more than 80% of the world's terrestrial biodiversity which include the plants, animals, fungi, and bacteria.

Generally, forest is also known as the lung of the Earth because the green plants in the forest absorb the carbon dioxide and release the oxygen into the atmosphere. Therefore, the forest is the gift that we should preserve from being destroyed. Malaysia is also one of the countries that has a lot of forest covering the land area. In this study, we are going to do an analysis about the forested area in Malaysia.

1.1 PURPOSE OF STUDY

The purpose of this study is to measure the change of the forest area in Malaysia. In this study, we are using one of the states in Malaysia situated in the south which is Johor. We are using the sample gathered of the forested area and also the non-forested area in Johor by using the unit of hectare.

1.2 INTEREST

We found this study is interesting because it will tell us about the forested area that has become one of the most valuable assets for our country. It is a home for hundreds of species including the flora and fauna and should be protected. We are hoping this study and analysis will provide the information and become a reference for the changes of the forested area in Johor.

1.3 EXPECTATION

The expectation for this study is to find whether there is a majority of changes in forested and non-forested areas in Johor as the year increases. By measuring the changes of the forested area, we can know whether the area of forest is increasing or decreasing by year. Therefore, we can plan the needed steps to preserve our forest including the habitat for hundreds of animals and plants from being destroyed.

2.0 DATASET

The dataset that we have chosen is a sample about forested and non-forested areas in Malaysia. From this data, we have calculated the difference of forested and non-forested areas(NF-F) for every year since 2000 until 2017.

NO. /n	YEAR /x	FORESTED (F) (hectare)	NON-FORESTED (NF) (hectare)	NF-F (hectare) /y
1	2000	474657	1423972	949315
2	2001	474657	1423972	949315
3	2002	474657	1423972	949315
4	2003	474657	1423972	949315
5	2004	472980	1425649	952669
6	2005	493072	1405557	912485
7	2006	508495	1390134	881639
8	2007	490209	1408420	918211
9	2008	441777	1456852	1015075
10	2009	468451	1430178	961727
11	2010	467262	1434338	967076
12	2011	466792	1434808	968016
13	2012	466792	1434808	968016
14	2013	466793	1434807	968014
15	2014	466768	1435225	968457
16	2015	449221	1452379	1003158
17	2016	449212	1452388	1003176
18	2017	447753	1468847	1021094

TABLE 1: Data of forested and non-forested areas.

From Table 1, we calculated mean, standard deviation and the variance. Calculation below is the steps used.

Mean = $\frac{\Sigma(NF-F)}{n}$ = 961448.5

Standard Deviation = $\frac{\sqrt{\Sigma(x-\bar{x})^2}}{n-1}$ = 35502.21

Variance = $(35502.21)^2$ = 1260406944

3.0 DATA ANALYSIS

3.1 HYPOTHESIS 1-SAMPLE TEST

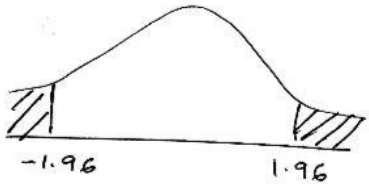
- Hypothesis
  H_0 : m = 960000
  H_1 : m ≠ 960000

- Calculate z-value
  $z = \frac{961448.5-960000}{(35502.21/\sqrt{18})} = 0.173100915$

- Calculate critical value
  α = 0.05,
  α/2 = 0.025,
  CV= $Z_{0.025}$=1.96

- Calculate p-value
  P(Z>0.173100915) = 0.43129

- Draw the graph



- Conclusion
  Since 0.43129 >0.025, we reject the null hypothesis H_0. Hence, the average difference between forested and non-forested areas should not equal 960000 hectare.

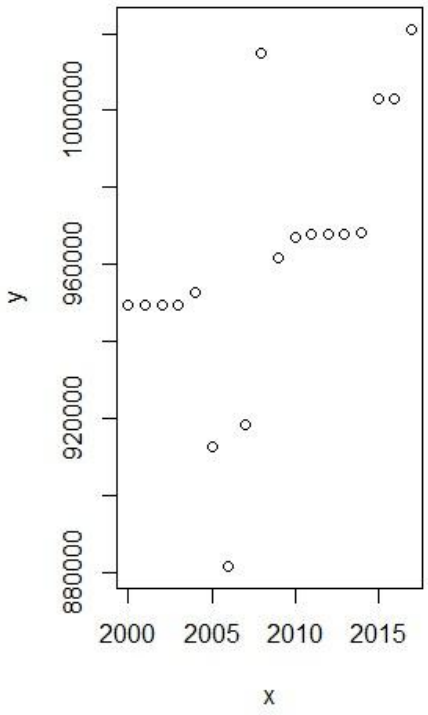
3.2 CORRELATION TEST

x	y	xy	x^2	y^2
2000	949315	1898630000	4000000	901198969225
2001	949315	1899579315	4004001	901198969225
2002	949315	1900528630	4008004	901198969225
2003	949315	1901477945	4012009	901198969225
2004	952669	1909148676	4016016	907578223561
2005	912485	1829532425	4020025	832628875225
2006	881639	1768567834	4024036	777287326321
2007	918211	1842849477	4028049	843111440521
2008	1015075	2038270600	4032064	1030377255625
2009	961727	1932109543	4036081	924918822529
2010	967076	1943822760	4040100	935235989776
2011	968016	1946680176	4044121	937054976256
2012	968016	1947648192	4048144	937054976256
2013	968014	1948612182	4052169	937051104196
2014	968457	1950472398	4056196	937908960849
2015	1003158	2021363370	4060225	1006325972964

2016	1003176	2022402816	4064256	1006362086976
2017	1021094	2059546598	4068289	1042632956836
Σ=36153	Σ=17306073	Σ=34761242937	Σ=72613785	Σ=16660324844791

TABLE 2: Calculation needed for correlation test.

- Scatter plot



- Hypothesis

  H_0: There is no correlation

  H_1: There is a correlation

- Calculate value of r

$$r = \frac{32761242937-(625666457169/18)}{\sqrt{[72613785-((36153)^2/18)][16660324844791-((17306073)^2/18)]}}$$

$$= 0.0000607829179670274$$

- Calculate the t-test

$$t = \frac{0.00006078291796702 74}{\sqrt{\frac{1-(0.00006078291796702 74)^2}{18-2}}}$$

$$= 0.000243$$

- Calculate critical value

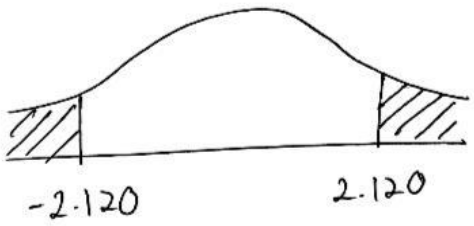
  $\alpha = 0.05$, df = 18-2 = 16

  Reject H_0 if:

  $t > t_{0.025,16} = 2.120$

  $t < -t_{0.025,16} = -2.120$

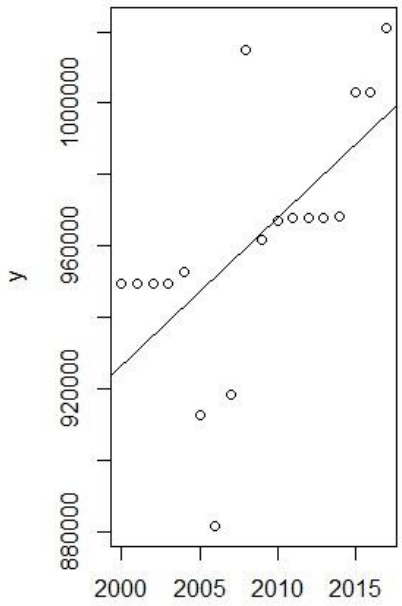
- Draw the graph



- Conclusion

  Since -2.120 < 0.000243 < 2.120, we fail to reject the null hypothesis H_0. Hence, there is no correlation between the year and the difference between forested and non-forested areas.

3.3 REGRESSION TEST



- Coefficients
  y-estimated = -7310158+4118x

- Intersection coefficient, b_0
  The negative y-terminal interception, b_0 = -7310158 for this regression model has no real meaning.

- Slope Coefficient, b_1
  b_1 = 4118. Thus, the average value of the difference between forested and non-forested areas increases by 4118 hectare on average, for each additional one year.

- Coefficient of Determination R^2
  H_0 : $R^2 = 0$ (no linear relationship)
  H_1 : $R^2 \neq 0$ (there is linear relationship)

  Multiple R^2 = 0.3835
  Adjusted R^2 = 0.345

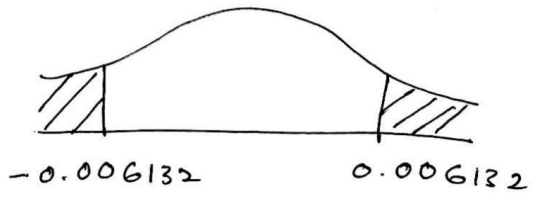
  Hence, there is a linear relationship between the number of years and the difference between forested and non-forested areas.

- Inference test
  H_0 : p = 0 (no linear relationship)

H_1 : p ≠ 0 (linear relationship exists)

Estimated Standard Error = 1305
t-value = 3.155
p-value = 0.006132

- Draw the graph



- Conclusion
  Since 3.155 > 0.006132, thus we reject the null hypothesis. The number of year affects the difference between forested and non-forested areas.

3.4 CHI SQUARE TEST (EQUAL FREQUENCIES/PROBABILITIES)

Year	Observed Difference
2000	949315
2001	949315
2002	949315
2003	949315
2004	952669
2005	912485
2006	881639
2007	918211
2008	1015075
2009	961727
2010	967076
2011	968016

2012	968016
2013	968014
2014	968457
2015	1003158
2016	1003176
2017	1021094

Table 3: Data about the year and observed difference.

- Hypothesis
  H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9 = p_{10} = p_{11} = p_{12} = p_{13} = p_{14} = p_{15} = p_{16} = p_{17} = p_{18}$
  H_1: At least one of the proportions is different from others.
- Calculate expected frequency
  Expected frequency, E $= \frac{(17306073)}{18} = 961448.5$

Year	Observed Difference	Expected Difference	$(O - E)^2/E$
2000	949315	961448.5	153.125
2001	949315	961448.5	153.125
2002	949315	961448.5	153.125
2003	949315	961448.5	153.125
2004	952669	961448.5	80.17031
2005	912485	961448.5	2493.555
2006	881639	961448.5	6624.958
2007	918211	961448.5	1944.443
2008	1015075	961448.5	2991.113
2009	961727	961448.5	0.080672

2010	967076	961448.5	32.93859
2011	968016	961448.5	44.86154
2012	968016	961448.5	44.86154
2013	968014	961448.5	44.83422
2014	968457	961448.5	51.08861
2015	1003158	961448.5	1809.439
2016	1003176	961448.5	1811.001
2017	1021094	961448.5	3700.235

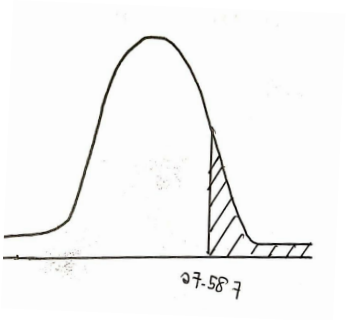
Table 4: Data needed to do the Chi-Square test.

Chi-Square test, $X^2 = \Sigma\frac{(O-E)^2}{E}$= 22286.07986

α = 0.05, df = 17

Reject H_0 if:

$X^2 > X^2_{17,0.05} = 27.587$



- Conclusion

  Since 22286.07986 > 27.587, we reject the null hypothesis H_0. Therefore, we reject the claim that the observed difference between forested and non-forested areas has equal proportions in the 18 years.

4.0 CONCLUSION

In finishing this project, we learned a lot of things that are related in the field of statistics and probability, especially inferential statistics. First of all, we learn that there is a lot of open

data source out there that we can use in order to do our own statistical test. From this, we have gained knowledge such as how to extract raw data from the dataset and how to do the related analysis in order to gather information that is needed to reach the conclusion for each test. For us, we think the most interesting findings are that there is no correlation between the year and the difference in forested and non-forested areas in Johor.

5.0 APPENDIX

https://www.dosm.gov.my/v1/index.php?r=column3/accordion&menu_id=amZNeW9vTXRydTFwTXAxSmdDL1J4dz09 -> Environment -> Forested and Non-Forested Areas, Malaysia, 2000 - 2017

Area	State	Year	Category	Hectares
Peninsular Malaysia	Johor	2000	Forested	474657
Peninsular Malaysia	Johor	2000	Non-Forested	1423972
Peninsular Malaysia	Johor	2001	Forested	474657
Peninsular Malaysia	Johor	2001	Non-Forested	1423972
Peninsular Malaysia	Johor	2002	Forested	474657
Peninsular Malaysia	Johor	2002	Non-Forested	1423972
Peninsular Malaysia	Johor	2003	Forested	474657
Peninsular Malaysia	Johor	2003	Non-Forested	1423972
Peninsular Malaysia	Johor	2004	Forested	472980
Peninsular Malaysia	Johor	2004	Non-Forested	1425649
Peninsular Malaysia	Johor	2005	Forested	493072
Peninsular Malaysia	Johor	2005	Non-Forested	1405557
Peninsular Malaysia	Johor	2006	Forested	508495
Peninsular Malaysia	Johor	2006	Non-Forested	1390134
Peninsular Malaysia	Johor	2007	Forested	490209
Peninsular Malaysia	Johor	2007	Non-Forested	1408420
Peninsular Malaysia	Johor	2008	Forested	441777
Peninsular Malaysia	Johor	2008	Non-Forested	1456852
Peninsular Malaysia	Johor	2009	Forested	468451
Peninsular Malaysia	Johor	2009	Non-Forested	1430178
Peninsular Malaysia	Johor	2010	Forested	467262
Peninsular Malaysia	Johor	2010	Non-Forested	1434338
Peninsular Malaysia	Johor	2011	Forested	466792
Peninsular Malaysia	Johor	2011	Non-Forested	1434808
Peninsular Malaysia	Johor	2012	Forested	466792
Peninsular Malaysia	Johor	2012	Non-Forested	1434808
Peninsular Malaysia	Johor	2013	Forested	466793
Peninsular Malaysia	Johor	2013	Non-Forested	1434807
Peninsular Malaysia	Johor	2014	Forested	466768
Peninsular Malaysia	Johor	2014	Non-Forested	1435225
Peninsular Malaysia	Johor	2015	Forested	449221
Peninsular Malaysia	Johor	2015	Non-Forested	1452379
Peninsular Malaysia	Johor	2016	Forested	449212
Peninsular Malaysia	Johor	2016	Non-Forested	1452388
Peninsular Malaysia	Johor	2017	Forested	447753
Peninsular Malaysia	Johor	2017	Non-Forested	1468847