UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143-01 KEBARANGKALIAN STATISTIK & ANALISIS DATA (PROBABILITY & STATISTICAL DATA ANALYSIS)

2021/2022

SEMESTER 2 PROJECT 2

Submitted to:

Dr Azurah A Samah

GROUP 3

NO.	NAME	STUDENT ID
1.	NEREA LANAI ANAK NEOL	A21EC0100
2.	YUSRA NADATUL ALYEEA BINTI YUSRAMIZAL	A21EC0151
3.	NUR AISYAH FATIHAH BINTI MOHAMED ROZI	A21EC0107
4.	LOKE RUI KEE	A20ET0329

Table of Contents

Introduction

In this given project, the dataset that has been chosen is a secondary data named Environment statistics 2021 which was acquired and collected by different government departments of Malaysia such as the Department of Environment, Ministry of Environment and Water, LAKU Management Sdn. Bhd. (Bintulu, Miri & Limbang), Sibu Water Board, Kuching Water Board, National Water Services Commission and Sarawak Rural Water Supply Department (JBALB). The dataset was also collated by the Department of Statistics Malaysia. The main objectives of these data are to show statistical tables for water quality of each state in Malaysia that affect the rivers. The most common standards used to monitor and assess water quality convey the health of ecosystems, safety of human contact, extent of water pollution and condition of drinking water. As we all know, water quality has a significant impact on water supply and oftentimes determines supply options for each state in Malaysia. Data that are featured in this research dataset is from the year 2016, 2017, 2018, 2019 and 2020.

This study is to justify which state has the highest scheduled waste, metered water consumption, production of water supplied and to compare Malaysia's water quality & the air pollutant emission. There is a variety of information we can extract from the data in this dataset. Among those, the one that we particularly interested in the study is about certain variables. For instance , the river water quality from the year 2018 to 2020, the scheduled waste by each state from the year 2019 to 2020, the production amount of water supplied by each state from the year 2016 to 2020 and the metered water consumption by each state & sector from the year 2016 or 2017.

The purposes of the study are as follows:

- To investigate whether the proportion of river water quality in Malaysia is the same between 2018 and 2020.
- To investigate whether there appears to be a linear relationship between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019.
- To determine whether there is a relationship between the number of river basins monitored and the polluted river water in 2018 & 2019 in Malaysia.
- To determine whether there appears to be a linear relationship between the use of domestic water in each state in 2016 and 2017.

The target population is every water supply & river in Malaysia and water supplied & used through the water meter by the National Water Services Commission. The studies are done through hypothesis testing, Chi-square testing (1 and 2 way contingency), correlation test and regression test with R-Studio. In the following part of the report, we will discuss data analysis and results then followed by discussion and conclusion along with the reference.

Data Analysis and Results

HYPOTHESIS TESTING (2-SAMPLE)

Test 1:

To investigate whether the proportion of river water quality in Malaysia is the same between 2018 and 2020 with a significance level of 0.05

Kategori/ *Category*	2018		2019		2020	
	Bil. *No.*	%	Bil. *No.*	%	Bil. *No.*	%
Bilangan lembangan sungai yang diawasi *Number of river basins monitored*	**143**	**100.0**	**144**	**100.0**	**144**	**100.0**
Tercemar/*Polluted*	10	7.0	10	7.0	7	4.9
Sederhana tercemar *Slightly polluted*	54	38.0	49	34.0	33	22.9
Bersih/*Clean*	79	55.0	85	59.0	104	72.2

Hypothesis Statement

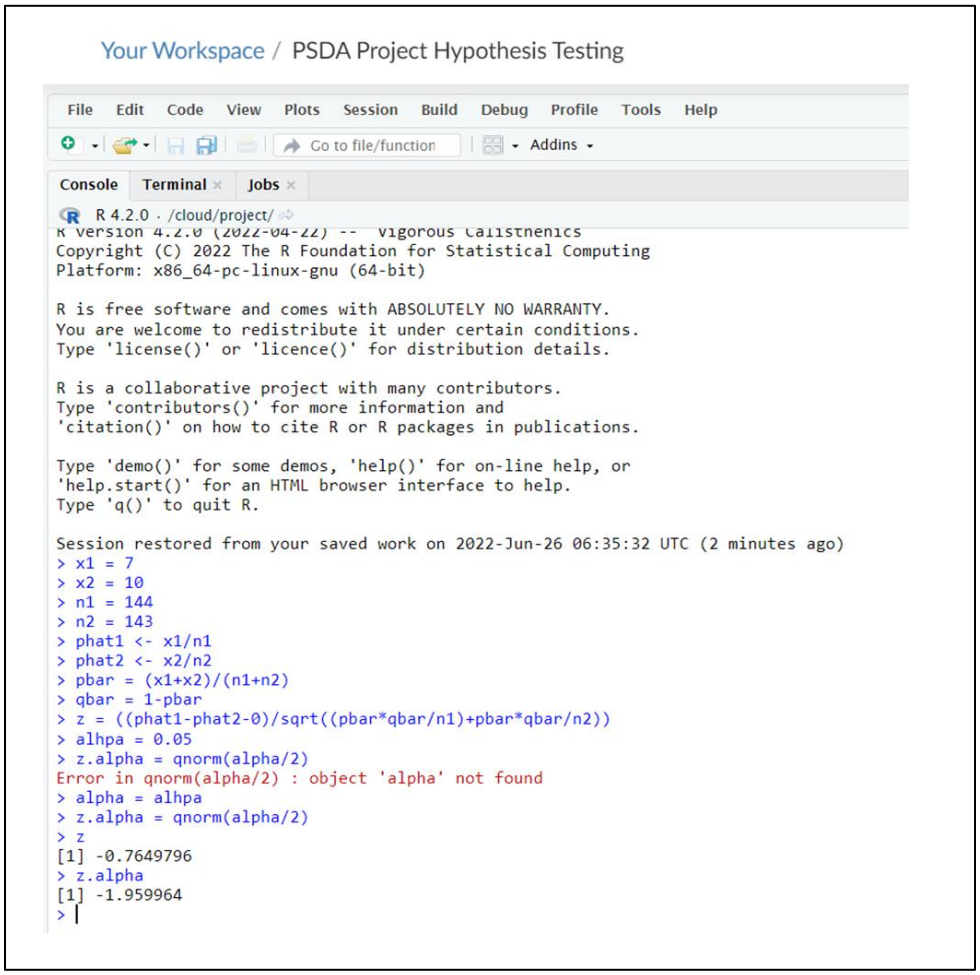
$H_0 : p1 = p2$

[Proportion of river water quality in Malaysia in 2020 = Proportion of river water quality in Malaysia in 2018]

$H_1 : p1 \neq p2$

[Proportion of river water quality in Malaysia in 2020 ≠ Proportion of river water quality in Malaysia in 2018]

Using R Studio:



Test Statistic

Total number of polluted rivers in 2020, x1: 7

Total number of polluted rivers in 2018, x2: 10

Total number of river basins monitored in 2020, n1: 144

Total number of river basins monitored in 2018, n2: 143

z = -0.7649796

Critical Value

$z_{0.025} = -1.959964$

Decision

H_0 is rejected since -0.7649796 > -1.959964

Conclusion

There is no sufficient evidence at 0.05 significance level to support that the proportion of river water quality in Malaysia in 2020 is the same as the river water quality in Malaysia in 2018.

CORRELATION

Test 2:

To investigate whether there appears to be a linear relationship between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019 at 5% significance level.

Jadual 14.4: Pengeluaran air yang dibekalkan mengikut negeri, 2016-2020

Table 14.4: Production of water supplied by state, 2016-2020

Juta liter sehari (JLH)
Million litres per day (MLD)

Negeri/ *State*	**2016**	**2017**	**2018**	**2019**	**2020**
Malaysia	**16,884**	**16,799**	**17,174**	**17,763**	
Johor	1,752	1,680	1,757	1,888	1,900
Kedah	1,370	1,370	1,446	1,474	1,487
Kelantan	475	475	491	503	507
Melaka	513	513	528	538	578
Negeri Sembilan	771	771	780	784	795
Pahang	1,109	1,109	1,158	1,194	1,269
Perak	1,314	1,314	1,333	1,357	1,392
Perlis	241	228	234	252	256
Pulau Pinang	1,058	1,058	1,073	1,093	1,108
Sabah	1,261	1,261	1,304	1,374	
Sarawak[1]	1,466	1,466	1,513	1,666	1,729
Selangor[2]	4,842	4,842	4,856	4,932	4,967
Terengganu	641	641	627	632	685
W.P. Labuan	71	71	74	76	82

Jadual 14.5: Penggunaan air bermeter mengikut sektor dan negeri, 2016-2020

Table 14.5: Metered water consumption by sector and state, 2016-2020

Juta liter sehari (JLH)
Million litres per day (MLD)

	Domestik/ *Domestic*					**Bukan Domestik/** *Non-Domestic*				
Negeri/ *State*	**2016**	**2017**	**2018**	**2019**	**2020**	**2016**	**2017**	**2018**	**2019**	**2020**
Malaysia	**6,501**	**6,454**	**6,653**	**6,824**	**7,430**	**4,240**	**4,332**	**4,469**	**4,716**	**4,076**
Johor	773	785	809	835	867	513	535	576	599	526
Kedah	525	523	544	551	574	200	195	201	208	193
Kelantan	163	162	164	163	170	76	79	85	85	75
Melaka	206	206	209	210	215	199	207	207	214	189
Negeri Sembilan	287	288	293	294	318	233	231	238	244	222
Pahang	342	333	333	339	367	238	250	256	288	264
Perak	655	648	665	679	706	262	260	268	278	266
Perlis	82	78	80	80	78	14	11	10	12	12
Pulau Pinang	492	482	491	499	533	335	344	349	343	314
Sabah	338	323	341	366	531	249	259	275	378	
Sarawak[1]	474	483	501	512	582	376	387	377	396	448
Selangor[2]	1,883	1,870	1,950	2,012	2,186	1,336	1,373	1,425	1,459	1,362
Terengganu	264	256	255	266	286	176	171	169	176	165
W.P. Labuan	17	17	18	18	17	33	31	33	36	40

Hypothesis Statement

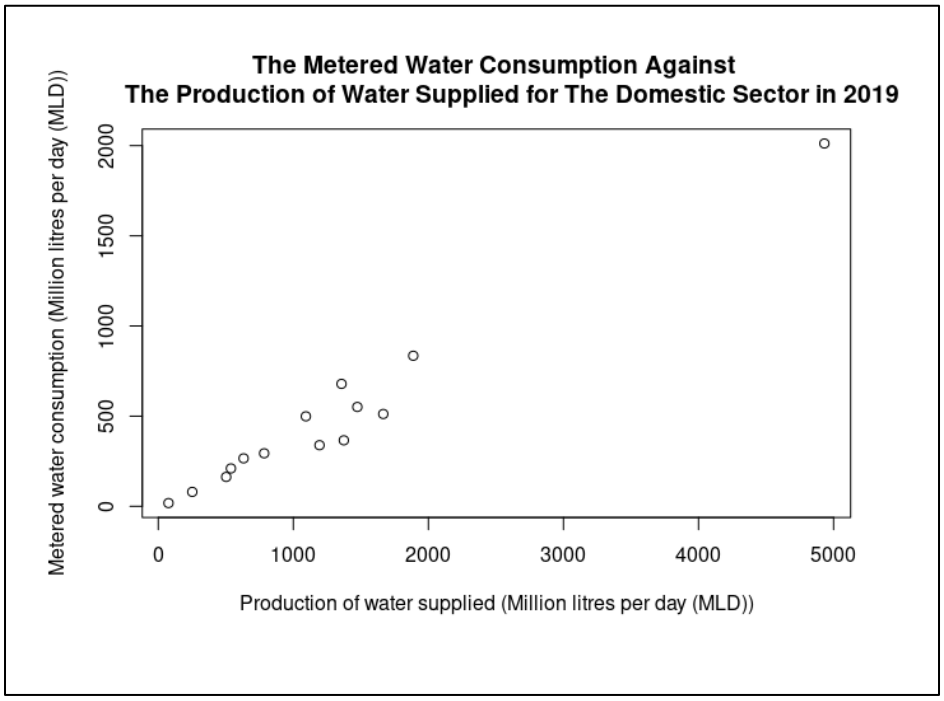
$H_0 : \rho = 0$

[There is no linear correlation between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019]

$H_1 : \rho \neq 0$

[There is a linear correlation between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019]

Plot


The Metered Water Consumption Against
The Production of Water Supplied for The Domestic Sector in 2019
Metered water consumption (Million litres per day (MLD))
0
500
1000
1500
2000
0
1000
2000
3000
4000
5000
Production of water supplied (Million litres per day (MLD))

Using R Studio:

```
> # 1. Set working directory
> setwd("/cloud/project")
>
> # 2. Load the library
> library("stats") # to run t-test
> library("psych") # to run descriptive analysis
> library("xlsx")
>
> # 3. Import the datafile into Rstudio
> data <- read_excel("Water Supplied and Consumption for Domestic 2019.xlsx")
>
> # 4. Test the assumptions
>
> # (a) Test for normality
> describe(data$Supplied)
   vars  n    mean      sd median trimmed   mad min  max range skew kurtosis     se
X1    1 14 1268.79 1185.66 1143.5 1062.92 766.5  76 4932  4856 1.95     3.66 316.88
> describe(data$Consumption)
   vars  n   mean     sd median trimmed    mad min  max range skew kurtosis     se
X1    1 14 487.43 494.46  352.5   399.5 258.71  18 2012  1994 1.98     3.59 132.15
>
> # 5. Run the analysis
> cor.test(data$Supplied, data$Consumption)

        Pearson's product-moment correlation

data:  data$Supplied and data$Consumption
t = 18.871, df = 12, p-value = 2.746e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9473919 0.9949308
sample estimates:
      cor
0.9835657

>
> # 6. Plot the data
> plot(Consumption ~ Supplied,
+      data = data,
+      main = "The Metered Water Consumption Against
+      The Production of Water Supplied for The Domestic Sector in 2019",
+      xlab = "Production of water supplied (Million litres per day (MLD))",
+      ylab = "Metered water consumption (Million litres per day (MLD))")
> cor(data$Supplied, data$Consumption)
[1] 0.9835657
> cor(data$Supplied, data$Consumption, use = "complete.obs")
[1] 0.9835657
```

Test Statistic

Pearson's correlation coefficient, r = 0.9835657

t = 18.871

Critical Value

Degree of freedom = 14 - 2 = 12

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

Decision

Reject H_0 since 18.871 > 2.179

Conclusion

There is sufficient evidence at 95% confidence level to support that there is a linear correlation between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019. Since the p-value is 0.0000000002746 < 0.05, there is a significant relationship between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019. The Pearson's correlation coefficient, r is 0.9835657 > 0.8, so it is a strong positive linear correlation. The higher the water supplied, the higher the metered water consumption for that particular state.

REGRESSION

Test 3:

To investigate whether there appears to be a linear relationship between the domestic metered water consumption in 2016 and the domestic meter consumption in 2017 in each state. Assume the confidence level to be 95% significance level, $\alpha = 0.05$.

Jadual 14.5: Penggunaan air bermeter mengikut sektor dan negeri, 2016-2020
Table 14.5: Metered water consumption by sector and state, 2016-2020

Juta liter sehari (JLH)
Million litres per day (MLD)

Negeri/ State	Domestik/ Domestic					Bukan Domestik/ Non-Domestic				
	2016	2017	2018	2019	2020	2016	2017	2018	2019	2020
Malaysia	**6,501**	**6,454**	**6,653**	**6,824**	**7,430**	**4,240**	**4,332**	**4,469**	**4,716**	**4,076**
Johor	773	785	809	835	867	513	535	576	599	526
Kedah	525	523	544	551	574	200	195	201	208	193
Kelantan	163	162	164	163	170	76	79	85	85	75
Melaka	206	206	209	210	215	199	207	207	214	189
Negeri Sembilan	287	288	293	294	318	233	231	238	244	222
Pahang	342	333	333	339	367	238	250	256	288	264
Perak	655	648	665	679	706	262	260	268	278	266
Perlis	82	78	80	80	78	14	11	10	12	12
Pulau Pinang	492	482	491	499	533	335	344	349	343	314
Sabah	338	323	341	366	531	249	259	275	378	313
Sarawak[1]	474	483	501	512	582	376	387	377	396	448
Selangor[2]	1,883	1,870	1,950	2,012	2,186	1,336	1,373	1,425	1,459	1,362
Terengganu	264	256	255	266	286	176	171	169	176	165
W.P. Labuan	17	17	18	18	17	33	31	33	36	40

Sumber/Source :
Kementerian Alam Sekitar dan Air
Ministry of Environment and Water

Suruhanjaya Perkhidmatan Air Negara
National Water Services Commission

Nota/Notes

[1] **Termasuk LAKU Management Sdn. Bhd. (Bintulu, Miri & Limbang), Lembaga Air Sibu, Lembaga Air Kuching dan Jabatan Bekalan Air Luar Bandar (JBALB) Sarawak**
Includes LAKU Management Sdn. Bhd. (Bintulu, Miri & Limbang), Sibu Water Board, Kuching Water Board and Sarawak Rural Water Supply Department (JBALB)

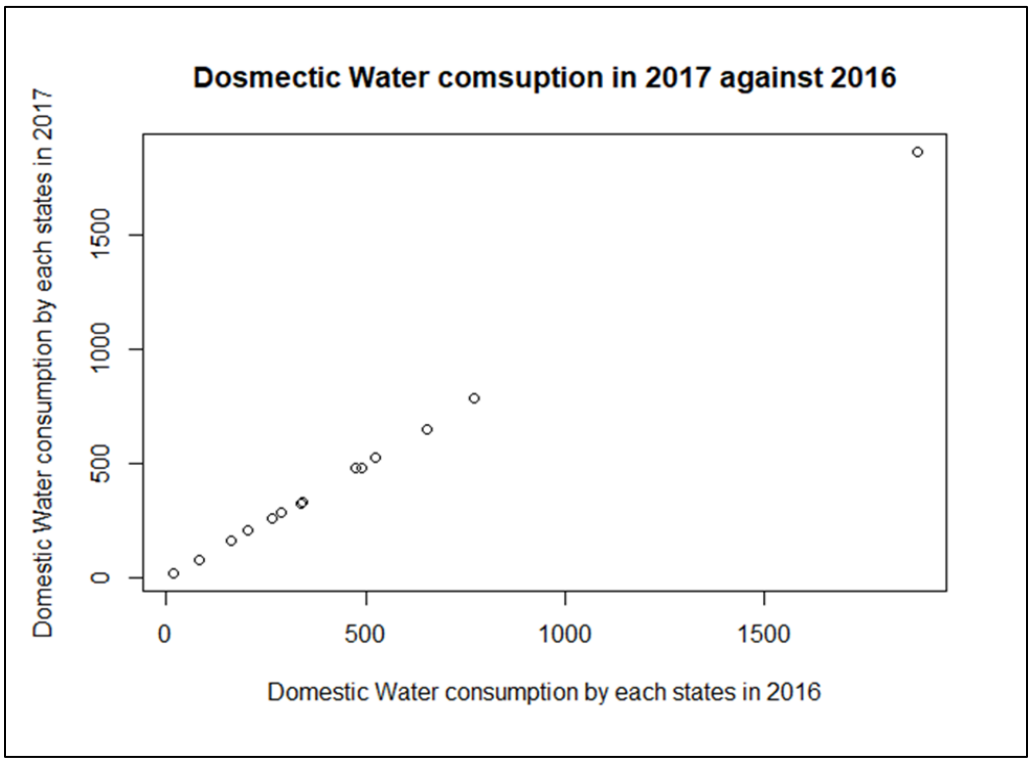
[2] **Termasuk W.P. Kuala Lumpur dan W.P. Putrajaya**
Includes W.P. Kuala Lumpur and W.P. Putrajaya

Hypothesis Statement

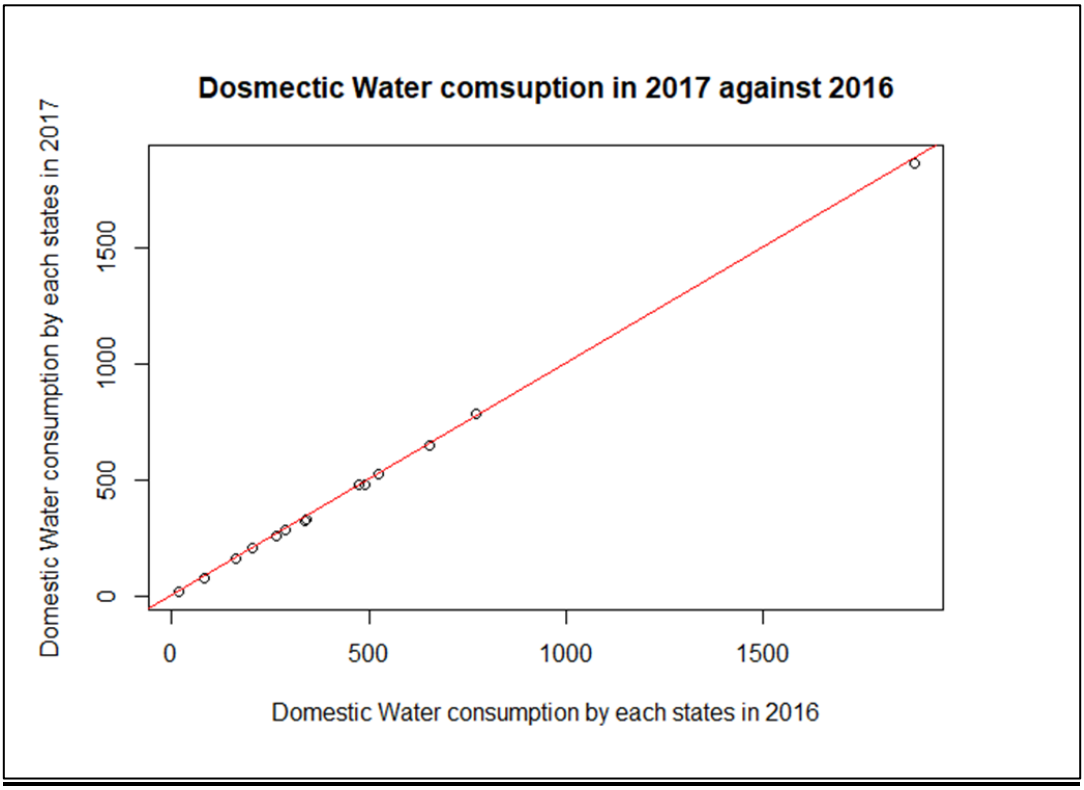
H0: $\beta 1 = 0$ (There is no linear regression between domestic water consumption by each state in 2016 with domestic water consumption by each state in 2017)

H1: $\beta 1 \neq 0$ (There is a linear regression between domestic water consumption by each state in 2016 with domestic water consumption by each state in 2017)

Scatter Plot



Scatter Plot with Regression Line



The independent variable (variable that used to explain the dependent variable) is domestic water consumption by each state in 2016 while the dependent variable) domestic water consumption by each state in 2017.

Using R Studio:

```
>water.consumption<-data.frame(year_2016=c(773,525,163,206,287,342,655,82,492,338,474,1883,264,17),year_2017=c(785,523,162,206,288,333,648,78,482,323,483,1870,256,17))
>water.consumption
year_2016 year_2017
1        773       785
2        525       523
3        163       162
4        206       206
5        287       288
6        342       333
7        655       648
8         82        78
9        492       482
10       338       323
11       474       483
12      1883      1870
13       264       256
14        17        17
>plot(water.consumption$year_2016,water.consumption$year_2017,main = 'Dosmectic Water consuption in 2017 against 2016',xlab = 'Domestic Water consumption by each states in 2016',ylab='Domestic Water
consumption by each states in 2017')
>water.regression<-lm(year_2016~year_2017, data=water.consumption)
>summary(water.regression)
Call:
  lm(formula = year_2016 ~ year_2017, data = water.consumption)
Residuals:
   Min      1Q  Median      3Q     Max
-16.4650  -2.3237   0.3088   5.2141  12.1147
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.780834   3.043092   0.585    0.569
year_2017   1.003419   0.004765 210.599   <2e-16 ***
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.881 on 12 degrees of freedom
Multiple R-squared:  0.9997,  Adjusted R-squared:  0.9997
F-statistic: 4.435e+04 on 1 and 12 DF,  p-value: < 2.2e-16
>abline(water.regression, col="red")
```

The r-squared value of regression, R^2 is 0.9997

This shows that is 99% of the domestic water consumption by each state in 2017.

is explained by domestic water consumption by each state in 2016.

$\alpha = 0.05$

Sample size, n = 14

Degree of freedom, df = 12

The regression line,

$\hat{y} = 1.781 + 1.003\ x$

Degree of freedom, df = 14 - 2

= 12

Critical Value

$-t_{0.025,12} = -2.179$

The shaded region is the rejection region.

$t_{0.025,12} = +2.179$

Negeri	(x)	(y)	y'	(y1-y')^2	x^2
Johor	773	785	777.1	62.41	597529
Kedah	525	523	528.356	28.687	275625
Kelantan	163	162	165.27	10.692	26569
Melaka	206	206	208.399	5.755	42436
Negeri Se	287	288	289.642	2.696	82369
Pahang	342	333	344.807	139.405	117649
Perak	655	648	658.746	115.477	429025
Perlis	82	78	84.027	36.325	6724
Pulau Pina	492	482	495.257	175.748	242064
Sabah	338	323	340.795	316.662	114244
Sarawak	474	483	477.203	33.605	224676
Selangor	1,883	1,870	1890.43	417.385	3545689
Terenggar	264	256	266.573	111.788	69696
W.P. Labu	17	17	18.832	3.356	289
SUM	6,501	6,454	6545.437	1459.991	5774584

Standard deviation of the variation of observations,

$$SSE = \sum (y - \hat{y})^2$$

SSE = 1459.991

$$s_\varepsilon = \sqrt{\frac{SSE}{n - k - 1}}$$

S^ε = 11.030

Standard error of the regression slope coefficient,

$$S_{b_1} = \frac{S_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{S_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Sb1: 4.764 x 10^-3

Test Statistic

Test statistic

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

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b_1} = standard error of the slope

= 210.599

Since the test statistic, t = 210.599> upper tail critical value, $t_{0.025,267}$ = 2.179 and it falls within the rejection region. Hence, we reject the null hypothesis. There is sufficient evidence that domestic water consumption by each state in 2016 affects the domestic water consumption by each state in 2017.

CHI - SQUARE INDEPENDENCE TEST (TWO-WAY CONTINGENCY)

Test 4:
To determine whether there is a relationship between the number of river basins monitored and the polluted river water in 2018 & 2019 in Malaysia at 5% significance level.

State the test hypothesis

H_0: No relationship between the number of river basins monitored and the polluted river water in 2018 & 2019.
H_1: Number of river basins monitored and the polluted river water in 2018 & 2019 have a relationship.

To find the critical value

Alpha: 0.05
df = (2-1) (2-1) = 1
$\chi^2_{1,0.05} = 3.841$

Using R Studio:

```
> #///// Test : Chi-square test of independence
> polluted <- c(10,19)
> clean <-c(79,85)
> d <- data.frame (polluted, clean)
> chisq.test(d, correct = FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 1.858, df = 1, p-value = 0.1729

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

Calculate the expected counts

Polluted river water	Number of river basins monitored				
	2018		2019		Total
	Obs.	Exp.	Obs.	Exp.	
polluted	10	9.7	19	2.1	20
clean	79	79.3	85	146.7	164
Total	89	89	95	95	184

Calculate the test statistic value

Cell, ij	Observed count, o_{ij}	Expected count, e_{ij}	$[o_{ij} - e_{ij}]^2 / e_{ij}$
1,1	10	9.7	0.0093
1,2	19	2.1	0.14
2,1	79	79.3	0.0011
2,2	85	146.7	0.26
		χ^2	0.1729

State the decision

Since the test statistic, χ^2 = 0.1729 is smaller than the critical value which is 3.841. It falls outside of the critical region. Besides that the p-value obtained is 0.1729 is bigger than 0.05. Hence, we failed to reject the null hypothesis.

Conclusion

There is enough evidence at 0.05 significance level to support that there is no relationship between the number of river basins monitored and the polluted river water in 2018 & 2019 and they are independent of each other.

Discussion & Conclusion

First of all, according to the first testing which is the 2-sample hypothesis testing to investigate whether the proportion of river water quality in Malaysia is the same between 2018 and 2020 with a significance level of 0.05. We found out that there is no sufficient evidence at 0.05 significance level to support that the proportion of river water quality in Malaysia in 2020 is the same as the river water quality in Malaysia in 2018. The result is kind a sad as the proportion is still the same between those 2 years but the polluted river water has decreased which is nice to see.

Besides, from the chi-square testing we found that the number of river basins monitored in 2018 & 2019 are not much different, which is 143 & 144 respectively. The number of polluted river water and the clean river water also increased in 2019. Meanwhile, the study also shows that there is no relationship between the number of river basins monitored and the polluted river water in 2018 & 2019 and they are independent of each other.

Next, during the first correlation test we found out that there is sufficient evidence at 95% confidence level to support that there is a strong positive linear correlation between the metered water consumption and the production of water supplied for the domestic sector in each state in 2019. Which then justify the hypothesis of the higher the water supplied, the higher the metered water consumption for that particular state. However, according to our regression test results, we observed that there is sufficient evidence that domestic water consumption by each state in 2016 affects the domestic water consumption by each state in 2017. The result is just as we already predicted since our rough assumption is that each year, the population of people will keep increasing in each state in Malaysia (more on states), there is a possibility that there is an increment in the usage of water.

REFERENCES

- Lecturer's note: Chapter 5
- Lecturer's note: Chapter 6
- Lecturer's note: Chapter 7
- Source of Data set: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=494&bul_id=dEVMYTl5VDNhQ2Roby9iK3VxOWM5dz09&menu_id=NWVEZGhEVlNMeitaMHNzK2htRU05dz09
- The Data Set : BMP_2020 A5_014 Jad_14 Alam Sekitar_Environment https://drive.google.com/drive/folders/1vIuu5jcMWYvIPuW11vVsvdwRPcsbGSFT?usp=sharing
- Correlation Test Between Two Variables in R. Retrieved from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
- Reading Data From Excel Files (xls | xlxs) into R. Retrieved from http://www.sthda.com/english/wiki/reading-data-from-excel-files-xls-xlsx-into-r
- Link e-portfolio
    - Loke Rui Kee: https://eportfolio.utm.my/user/loke-rui-kee/psda-project-reflection
    - Nerea Lanai: https://eportfolio.utm.my/artefact/artefact.php?artefact=1393471&view=315991
    - Nur Aisyah Fatihah: SECI2143-01 KEBARANGKALIAN STATISTIK & ANALISIS DATA (PROBABILITY & STATISTICAL DATA ANALYSIS) - MyePortfolio@UTM
    - Yusra Nadatul Alyeea: https://eportfolio.utm.my/user/yusra-nadatul-alyeea-binti-yus/seci2143-01-kebarangkalian-statistik-analisis-data-probability-statistical-data-analysis