



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SEMESTER 2, SESSION 2021/2022

SCHOOL OF COMPUTING

SECI2143_SEC 03
(PROBABILITY & STATISTICAL DATA ANALYSIS)

PROJECT 2_GROUP 3

PREPARED FOR:
LECTURER: DR ROZILAWATI DOLLAH @ MD ZAIN

PREPARED BY:

GROUP NAME	NAME	MATRIC NO
UTM SOUL	GOH JUN BOON	A21EC0179
	LEE JUN KANG	A21EC0194
	SANG YEN TING	A21EC0225
	TAN LI SIN	A21EC0231

Table of Content

No.	Content	Page
1	Introduction	2
2	Dataset	2
3	Test 1: One hypothesis Sample	3-4
4	Test 2: Correlation	5-6
5	Test 3: Regression	7-8
6	Test 4: Chi-square of Independence	9-10
7	Discussion	11
8	Conclusion	11
9	Appendix	11

Introduction

In this Project 2, we are required to conduct an inference statistical analysis that covers the following analysis such as the Hypothesis testing 1-sample or 2-sample, correlation, and regression which are compulsory must include in our Project. Next, we must also conduct at least one inference statistical analysis, choosing from Goodness of fit test, Chi Square test of independence, and ANOVA.

We had chosen the data set from the Kaggle web. The topic of the dataset we chose is Student Flexibility in Online Learning. The dataset consists of various variables such as Education Level, Institution Type, Gender, Age, Device, IT Student, Location, Financial Condition, Internet Type and Network Type.

The purpose we choose this dataset for our project is because it contains distinct variables that are suitable for us to conduct statistical analysis. For example, we get to test the mean of Student's age by using one sample test. Next, we can measure the strength of the relationship between student age and the type of device students use by using correlation. Furthermore, we are able to measure the relationship between financial health and institution type by making use of regression. Lastly, we also get to know the significant relationship between internet type and network type by implementing the Chi Square test.

Dataset

The dataset we choose from Kaggle web. The topic of the dataset is Student Flexibility in Online Learning.

The dataset consists of various variables and their parameters such as Education Level (School/University/Others), Institution Type (Private/Public), Gender (Male/Female), Age (9-27), Device (Computer/Mobile/Tablet), IT Student (True/False), Location (Town/Rural), Financial Condition (Mid/Poor/Others), Internet Type (Mobile Data/Wifi) and Network Type (3G/4G/Others).

We do the data pre-processing with calculation and ensure that the expected count ≥ 5 , otherwise the data analysis will be inaccurate for the chi-square test of independence.

We choose the variables as it is clear and fulfill our requirements of the dataset. The possible outcomes of our analysis is there might be sufficient evidence that the sample is from a group of students with an average age greater than 17 years for hypothesis sample test. For correlation, we would like to find whether there is sufficient evidence to prove that there is a weak linear relationship between the student age and the types of device used by students. We might know that it is very low to prove that financial status affects the type of institution. There might be sufficient evidence that internet type and network type are not independent.

Test 1: One hypothesis sample test

According to the data of 'Student Flexibility in Online Learning', the mean of a student's age with a different education level is 17 years old. This is a one-sample test to test the claim that the sample is from a population of Student's age with a mean greater than 17 years old. Assume the confidence level to be 95%, significant level, $\alpha = 0.05$. Let the population mean of the student's age be μ .

$$H_0: \mu = 17$$

$$H_1: \mu > 17$$

$$\alpha = 0.05$$

Sample size, $n = 456$

Sample mean, $\bar{X}_{\text{bar}} = 23$

Sample Standard Deviation = 2.08048948

Test Statistics,

$$Z = \frac{\bar{X}_{\text{bar}} - \mu}{S/\sqrt{n}}$$

$$Z = \frac{23 - 17}{2.08048948/\sqrt{456}}$$

$$= 61.58$$

Critical value , $c.v = z_{0.05} = 1.644854 = 1.645$

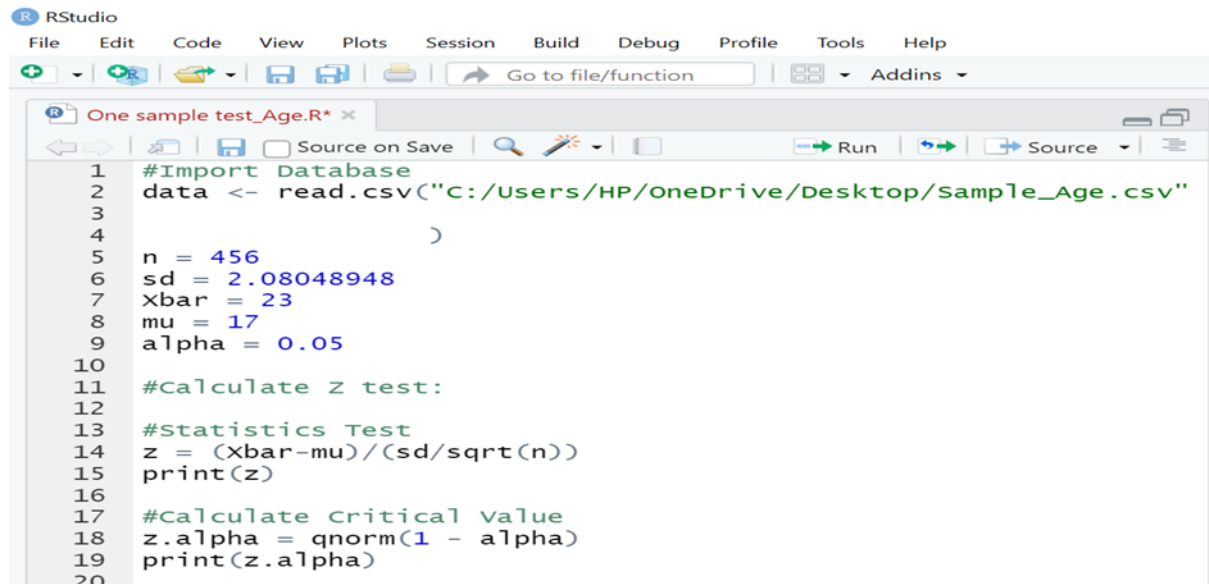
Decision: Since, $Z = 61.58$ greater than 1.645 falls within the critical region, we reject the null hypotheses, H_0 .

Conclusion: We have sufficient evidence to claim that the sample is from a population of Student's age with a mean greater than 17 years old.

**** R programming calculations are shown here ****

First, we need to import our dataset

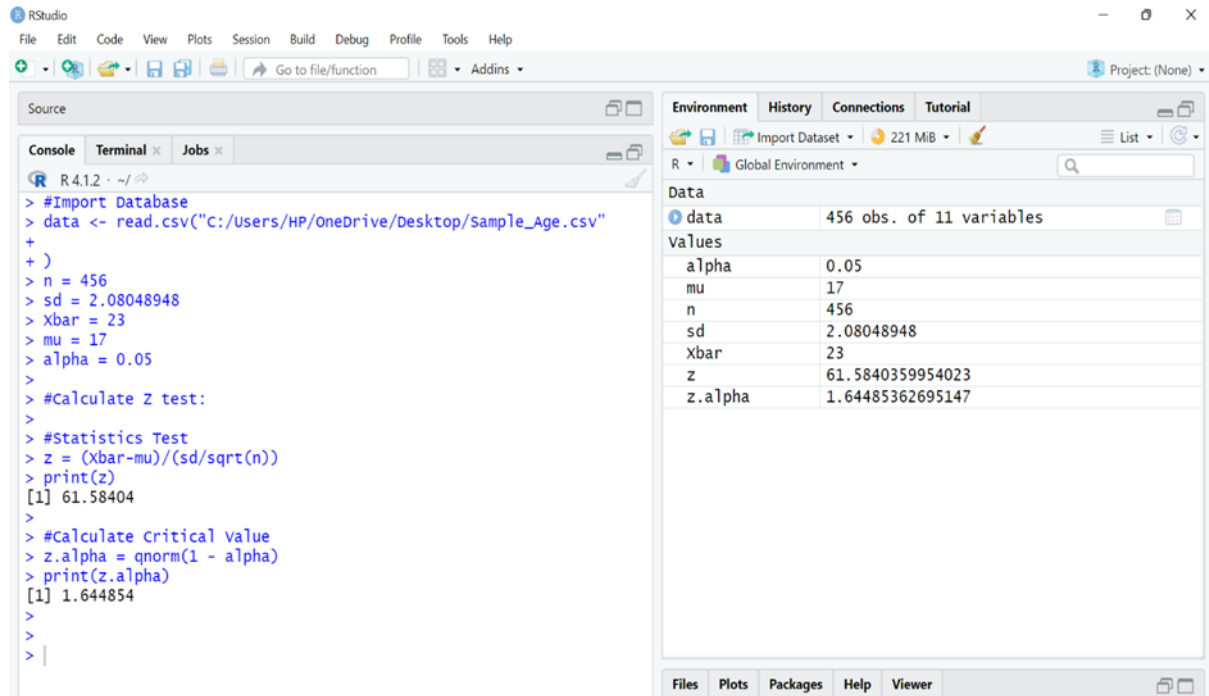
& Enter our R script:



The screenshot shows the RStudio interface with a script editor open. The script is titled 'One sample test_Age.R*' and contains the following R code:

```
1 #Import Database
2 data <- read.csv("C:/Users/HP/OneDrive/Desktop/Sample_Age.csv"
3
4 )
5 n = 456
6 sd = 2.08048948
7 xbar = 23
8 mu = 17
9 alpha = 0.05
10
11 #Calculate z test:
12
13 #Statistics Test
14 z = (xbar-mu)/(sd/sqrt(n))
15 print(z)
16
17 #Calculate Critical Value
18 z.alpha = qnorm(1 - alpha)
19 print(z.alpha)
20
```

Then, Console Will Display:



The screenshot shows the RStudio interface with the console and environment panes. The console displays the output of the R script, and the environment pane shows the variables created during the execution.

Console Output:

```
> #Import Database
> data <- read.csv("C:/Users/HP/OneDrive/Desktop/Sample_Age.csv"
+ )
+ )
> n = 456
> sd = 2.08048948
> xbar = 23
> mu = 17
> alpha = 0.05
>
> #Calculate z test:
>
> #Statistics Test
> z = (xbar-mu)/(sd/sqrt(n))
> print(z)
[1] 61.58404
>
> #Calculate Critical Value
> z.alpha = qnorm(1 - alpha)
> print(z.alpha)
[1] 1.644854
>
>
> |
```

Environment Pane:

Variable	Value
alpha	0.05
mu	17
n	456
sd	2.08048948
xbar	23
z	61.5840359954023
z.alpha	1.64485362695147

Test 2: Correlation Analysis to investigate the relationship between the student age and the types of device used by students.

This test is to measure the strength of the relationship between the student age and the types of device used by students.

Assume the confidence level to be 95%, significant level, $\alpha = 0.05$.

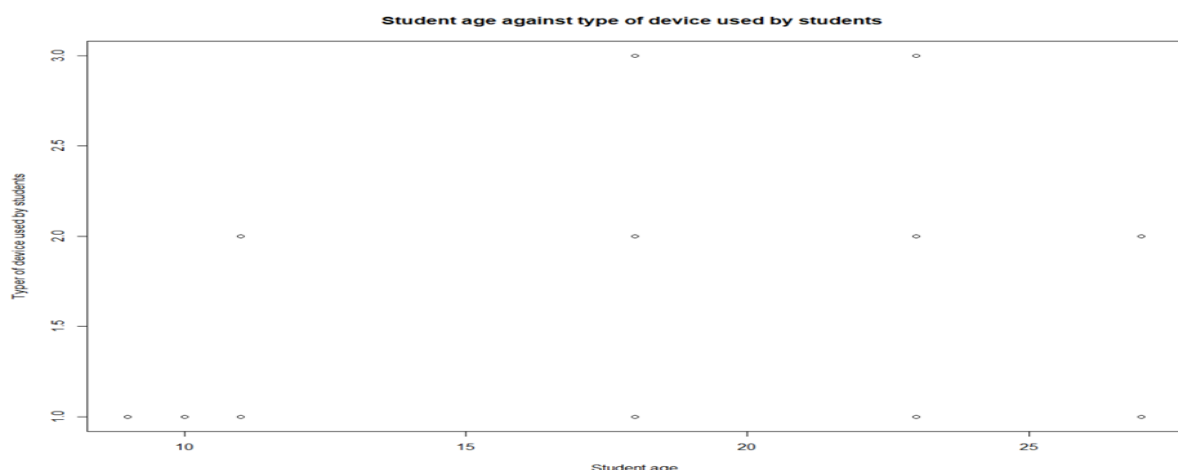
H0: $\rho = 0$ (no linear correlation between the student age and the types of device used by students.)

H1: $\rho \neq 0$ (linear correlation exists between the student age and the types of device used by students.)

```

1 setwd("C:/Programming/PSDA project/")
2 getwd()
3
4 library(dplyr)
5
6 students_adaptability_level_online_education <- readxl::read_excel("students_adaptability_level_online_education.xlsx")
7
8 #set
9 #mobile = 1
10 #computer = 2
11 #tab = 3
12
13 students_adaptability_level_online_education <- students_adaptability_level_online_education %>% mutate(device = replace(
14 students_adaptability_level_online_education <- students_adaptability_level_online_education %>% mutate(device = replace(
15 students_adaptability_level_online_education <- students_adaptability_level_online_education %>% mutate(device = replace(
16
17 students_adaptability_level_online_education$device <- as.numeric(students_adaptability_level_online_education$device)
18
19 x <- students_adaptability_level_online_education$age
20 y <- students_adaptability_level_online_education$device
21
22 cor(x,y)
23
24 plot(x,y,main = "student age against type of device used by students",xlab = "student age",ylab = "type of device used by students")
25 cor.test(x,y)
26
27

```



$$\alpha = 0.05$$

$$\alpha = 0.0 \text{ as it is 2 tail test}$$

Using RStudio

$$\text{Correlation coefficient, } r = 0.3512713$$

$$\text{Degree of freedom, } df = 1203$$

$$\text{Test statistic, } t = 13.013$$

Critical value,

$$-t(0.025, 1203) = -1.968$$

$$t(0.025, 1203) = 1.968$$

Decision: Since the test statistic, $t = 13.013$ is not in between $-t(0.025, 1203) = -1.968$ and $t(0.025, 1203) = 1.968$. it fall within the rejection region. Hence we success to reject the null hypothesis.

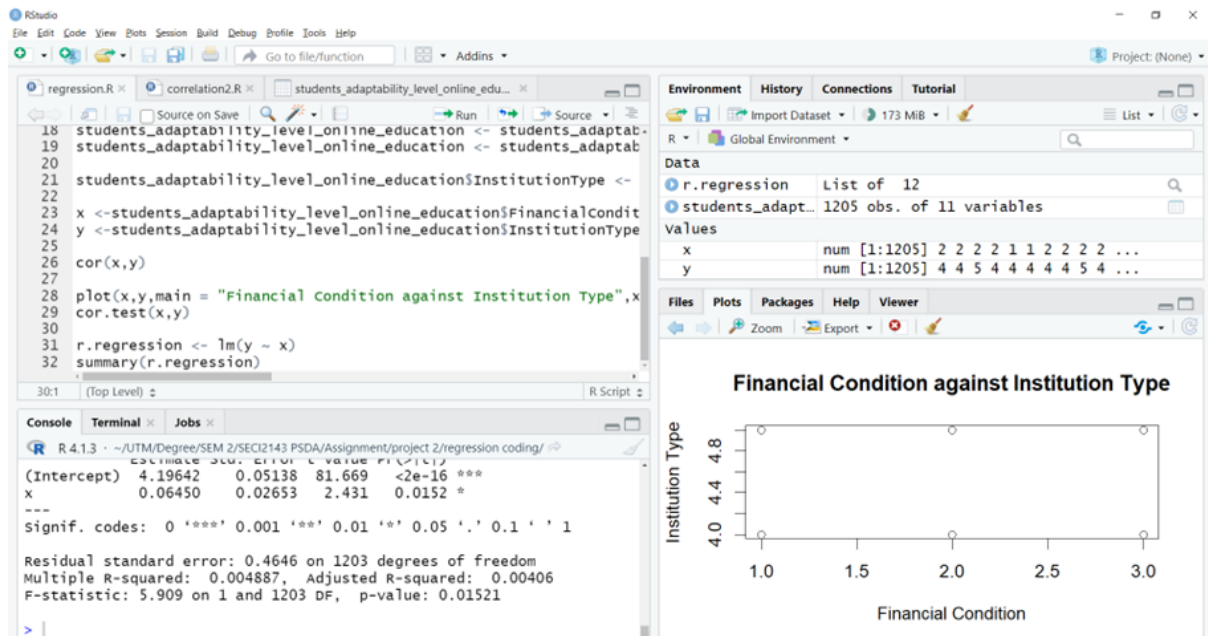
Conclusion: Since the correlation coefficient, $r = 0.3512713$ which is positive and falls within 0 and 0.5, hence it has a weak linear relationship between the student age and the types of device used by students. There is sufficient evidence to prove that there is a weak linear relationship between the student age and the types of device used by students.

Test 3: Regression analysis between Financial condition and institution type.

This test is to measure the relationship between financial condition and institution type.

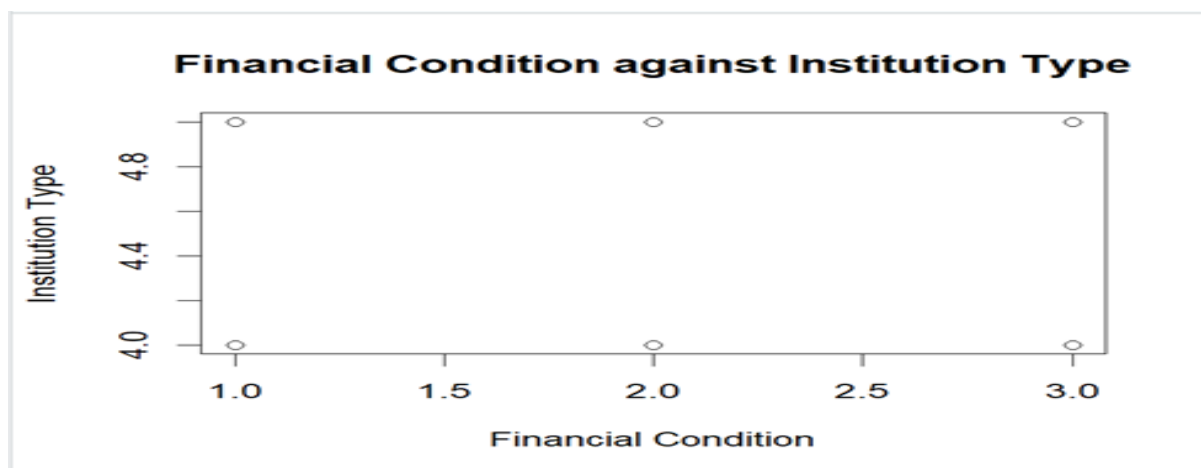
Assume that the confidence level = 95%, significance level, $\alpha = 0.05$.

The diagram below is the Rstudio Programming



The null hypothesis will be: $H_0: p=0$ (the financial condition is independent of the institution type.)

The alternative hypothesis will be: $H_a: p \neq 0$ (the financial condition is not independent of and the institution type.)




```

call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.3899 -0.3254 -0.3254  0.6746  0.7391

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19642    0.05138   81.669  <2e-16 ***
x            0.06450    0.02653    2.431   0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4646 on 1203 degrees of freedom
Multiple R-squared:  0.004887, Adjusted R-squared:  0.00406
F-statistic: 5.909 on 1 and 1203 DF, p-value: 0.01521

```

For the table we know that the intercept is 4.19642, while the slope is = 0.0645

And for the p-value is = 0.0152 and the t value is = 2.431

And the value of R^2 is = 0.00488

We know that when $\alpha = 0.05$. And CI= 95%, from the distribution table we will know that the critical value = -1.968 and 1.968.

Decision: Since the test statistic, $t = 2.431$ is greater than 1.968 and it fail in the rejection region, therefore we reject the null hypothesis.

Conclusion: Since the R^2 is = 0.00488, it means that financial condition can explain only 0.48% of the variation in the institution type. Which is very low to proof that the financial condition will affect the institution type.

Test 4: Chi-Square test of independence to determine whether there is a significant relationship between internet type and network type.

Assume the confidence level to be 95%, significant level, $\alpha = 0.05$.

H_0 : The internet type is independent of the network type.

H_1 : The internet type is not independent of the network type.

**** Manual calculations are shown here ****

Network Type	Internet Type	
	Mobile Data	Wifi
2G	14	5
3G	334	77
4G	347	428

Network Type	Internet Type				Total
	Mobile Data		Wifi		
	Obs.	Exp.	Obs.	Exp.	
2G	14	10.9585	5	8.0415	19
3G	334	237.0498	77	173.9502	411
4G	347	446.9917	428	328.0083	775
Total	695		510		1205

Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	$[o_{ij} - e_{ij}]^2 / e_{ij}$
1, 1	14	10.9585	0.8442
1, 2	5	8.0415	1.1504
2, 1	334	237.0498	39.6513
2, 2	77	173.9502	54.0347
3, 1	347	446.9917	22.3681
3, 2	428	328.0083	30.4820
χ^2			148.5307

**** R programming calculations are shown here (output will display) ****

```
Mobile Data Wifi
2G      14    5
3G     334   77
4G     347  428
> # perform chi-square test on the data table
> chisq.test(tbl, correct=FALSE)

Pearson's Chi-squared test

data:  tbl
X-squared = 148.53, df = 2, p-value < 2.2e-16
> #critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=2, lower.tail=FALSE)
> print(x2.alpha)
[1] 5.991465
```

← **contingency table**

degree of freedom

p-value

test statistic

critical value

Decision: Since the test statistic, $X^2 = 148.53$ is greater than the critical value which is 5.991465. It falls within the critical region. Besides that, the p-value obtained is $2.2e-16$ is smaller than 0.05. Hence, we reject the null hypothesis.

Conclusion: Since $148.53 > 5.991465$, thus reject H_0 .

There is sufficient evidence that the internet type and the network type are not independent.

Discussion

The average age of students at different levels of education is 17. This is a one-sample test to test the claim that the sample is from a group of students with an average age greater than 17 years. Measures the strength of the relationship between student age and the type of device students use. Measures the relationship between financial health and institution type. Measures the significant relationship between internet type and network type.

Conclusion

What have you learned from all activities done in Project 2 (choosing dataset, pre-processing and analysis process etc.)

We learned to choose datasets because the topics we want to choose must be live datasets and topical. In preprocessing, we need to know in advance which data is suitable for the association between variables and need to apply these tests to prove the association between variables. Of course, the variables we choose in preprocessing can affect subsequent analyses. Therefore, it is necessary to predict in advance whether the conditions for each test will be met. For example, a chi-square test of independence must ensure that the expected count ≥ 5 , otherwise the data analysis will be inaccurate. When these problems are found in preprocessing, we can avoid this by choosing a better dataset. After the preprocessing is over, we will enter the analysis process. As we analyze the process, we address our tests with reference to what we have learned in the classroom. We use the Rstudio method because it can quickly calculate the data content of the test and analyze the results super fast. After the analysis is done, we will know the conclusion relationship we want and the final result.

What is your best/interesting findings from your results?

In a sample test, we have sufficient evidence that the sample is from a group of students with an average age greater than 17 years. In addition to the correlation analysis, there is sufficient evidence for a weak linear relationship between student age and the type of device students use. In addition to this, we also used regression analysis to measure the relationship between financial health and institution type. We know that it is very low to prove that financial status affects the type of institution. Finally, we examined the relationship between Internet type and network type using the independence chi-square, and we obtained sufficient evidence that internet type and network type are not independent.

We can admit that most online learning is now over the age of 17. Students of different ages use different types of devices, as there is no rule about which device is used more by any age group. And the financial status of the students does not affect the type of school, so you are admitted to the school based on your strength, not because you have money. We will also know that internet type and network type are interdependent, so different internet types affect the strength of the network.

Appendix

Source of data:

<https://www.kaggle.com/datasets/shariful07/student-flexibility-in-online-learning>