UTM UNIVERSITI TEKNOLOGI MALAYSIA SCHOOL OF COMPUTING
Faculty of Engineering

UNDERGRADUATE UTM
SCHOOL OF COMPUTING

GROUP ASSIGNMENT
JUN 2021

SECI 2143 STATISTICS AND PROBABILITY

PROJECT 2

SECTION:
02

GROUP:
MYR

NO.	Name	Matrix
1.	Maizatul Afrina Safiah Binti Saiful Azwan	A20EC0204
2.	Madihah binti Che Zabri	A20EC0074
3.	Yong Zhi Yan	A20EC0172
4.	Kong Jia Rou	A20EC0198

LECTURER’S NAME:
DR CHAN WENG HOWE

SUBMISSION DATE:
3rd JULY 2021

Table of Content

Introduction

Examination is an assessment which all students in the world need to face. Getting a good grade in examinations is all that students hoped for. Therefore, their performance in the examination hall is very important. They must be in good spirits to perform well in examinations. It is claimed that there are many factors that may affect the students' performance in examinations (Mushtaq & Khan, 2012). One of the factors is the lunch. We questioned the relationship between the lunch factor and the students' performance in examinations. Apart from that, we also questioned the relationship between the students' ability in two different subjects. Do the students who are good in math are also good in writing? Do the students who are good in writing are also good in reading ?

1.1 Purpose

The purpose of the project :

- To investigate the relationship between the students' math score and the students' writing score.
- To investigate the relationship between the students' writing score and the students' reading score
- To investigate the relationship between the lunch and the students' performance in examinations.

1.2 Expectation

It is predicted that the lunch will affect the students' performance in examinations. We also predict that the students who are good in math are also good in writing. We expect that the students who are good in writing are also good in reading.

Data Set

2.1 Data Pre-processing

Secondary data from the website (students-performance-in-exams) prepared by Jakki Sheshapanpu is used in this project. The inspiration of Sheshapanpu's investigation is to understand the influence of the parents' background, test preparation etc on students' performance. The targeted population is the students of public high schools in the United State. In our project, the data is analyzed using several tests : hypothesis sample test, correlation test, regression test, chi-square test .

2.2 Parameter

Parameters included in our project are Lunch, Math score, Reading score, Writing score and Result. These parameters are used to carry out the tests.

2.3 Prediction on Outcome of Test

In our prediction, we predict that the population mean of average score is 50.00. In correlation test, we predict there is a relationship between the mathematics scores and the writing scores among the students. While in regression test, we predict there is a relationship between writing scores and reading scores of the students. In the chi-squared test, we also predict that there is a relationship between there is a relationship between the lunch and the students' performance in examinations.

Data Analysis

3.1 Test

3.1.1 Hypothesis Sample Test

By using the average score (avgr) in the dataset, we want to test whether the population mean for the average score is equal to 50 (the passing score for the exam).

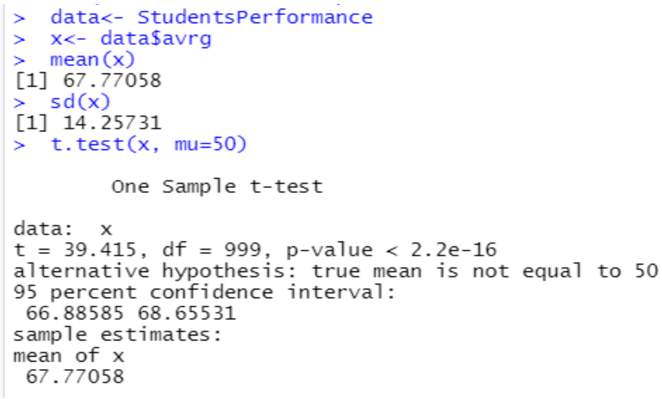
Let μ to be the population mean of the sample, H_o is the null hypothesis where the population mean is equal to 50 and H_1 is the alternative hypothesis where the population mean is more than 50.

H_o: $\mu = 50.00$

H_1: $\mu > 50.00$

This hypothesis statement will be tested with a significance level of 0.05 where $\alpha = 0.05$.

As the population variance is unknown and the sample size is relatively large ($n = 1000$), hence we use t-test to test the hypothesis.



```
>  data<- StudentsPerformance
>  x<- data$avrg
>  mean(x)
[1] 67.77058
>  sd(x)
[1] 14.25731
>  t.test(x, mu=50)

        One Sample t-test

data:  x
t = 39.415, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 66.88585 68.65531
sample estimates:
mean of x 
 67.77058 
```

Figure 1 : Hypothesis sample test result of RStudio

The critical values are ± 1.962. The p-value of 2.2e-16 is less than the significance level of α = 0.05. The test statistic t = 39.415 falls in the critical region bounded by ± 1.962. The claimed mean of 50 does not fall within the confidence interval of 66.88585 < μ < 68.65531. Hence, there is sufficient evidence to reject the null hypothesis. Thus, we can conclude that the overall performance of students in this examination is higher than the passing line, and the alternative hypothesis is accepted.

3.1.2 Correlation Test

The purpose of this correlation test is we want to know if there is any relationship between the students' mathematics scores and the students writing scores.

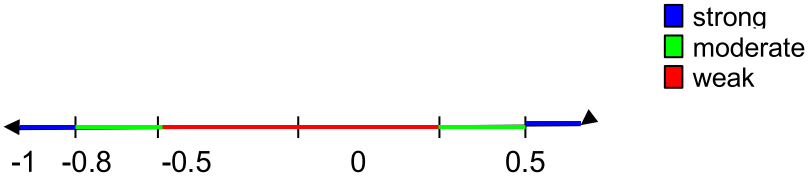
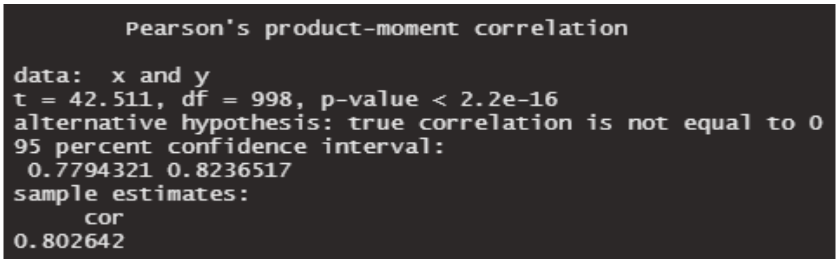


Figure 2: The strength of the relationship's indicator

From figure 2, we can know that the closer the r's value to 1, the stronger the positive linear relationship while the closer the r's value to -1, the stronger the negative linear relationship. However, the linear relationship will become weaker if the value of r is close to 0.



```
        Pearson's product-moment correlation

data:  x and y
t = 42.511, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7794321 0.8236517
sample estimates:
     cor
0.802642
```

Figure 3: Correlation test of RStudio

Figure 3 shows the result of correlation coefficient (r) using Pearson's method which is 0.802642. We use Pearson's method as both of the variables are ratio-type data. Based on the value of r, it shows that there is a strong relationship between the students' mathematic scores and the students' writing score. Also, the positive value of r indicates that the values of mathematics scores and writing scores increase together.

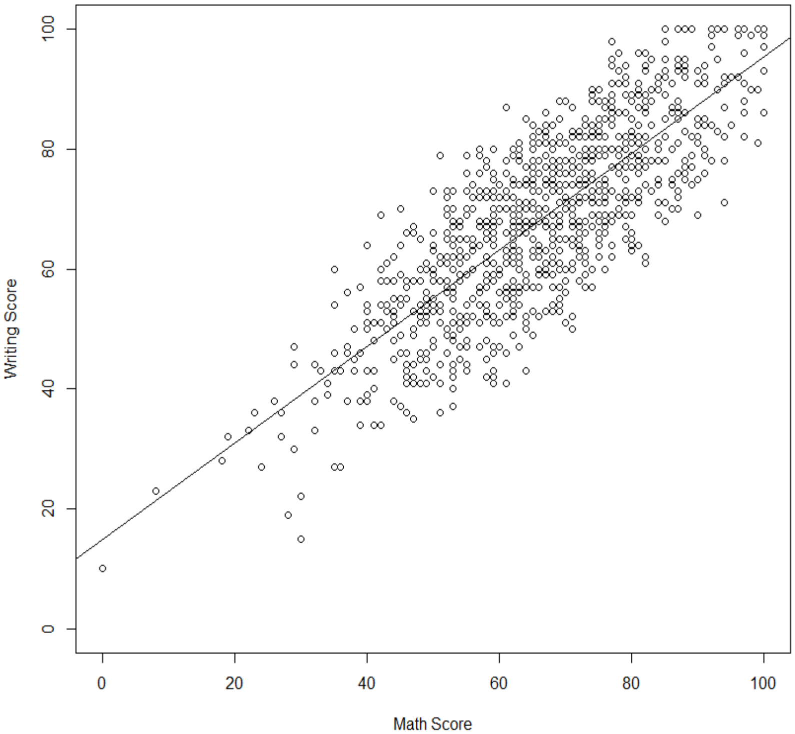
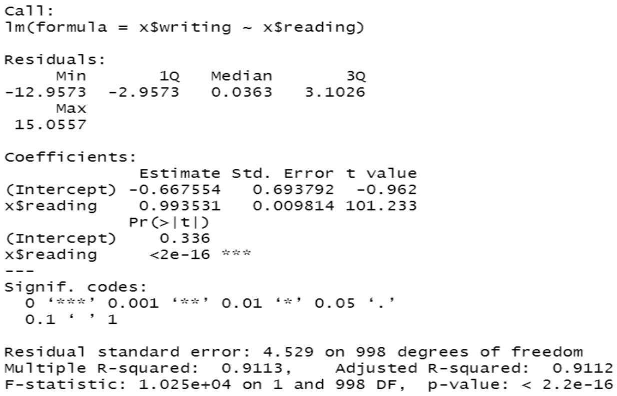


Figure 4: Scatter plot of Writing Scores against Math Scores with regression line

Based on the graph above, we are using two variables which are math score as x-axis and writing score as y-axis. It is clearly shown that both of the variables have positive linear relationships and strong relationship. In general, it appears that students with higher mathematics scores tend to have higher writing scores. Thus, we can conclude that this test provides sufficient evidence to support that there is a relationship between the mathematics scores and the writing scores among the students.

3.1.3 Regression Test

By doing this regression test, we are able to analyse if there is any relationship between writing scores and reading scores of the students.
The dependent variable of this relationship is Reading while Writing will be the independent variable.



```
Call:
lm(formula = x$writing ~ x$reading)

Residuals:
     Min       1Q   Median       3Q 
-12.9573  -2.9573   0.0363   3.1026 
     Max 
 15.0557 

Coefficients:
             Estimate Std. Error t value
(Intercept) -0.667554   0.693792  -0.962
x$reading    0.993531   0.009814 101.233
            Pr(>|t|)    
(Intercept)    0.336    
x$reading     <2e-16 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 4.529 on 998 degrees of freedom
Multiple R-squared:  0.9113,	Adjusted R-squared:  0.9112 
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

Figure 5: Regression test result of RStudio

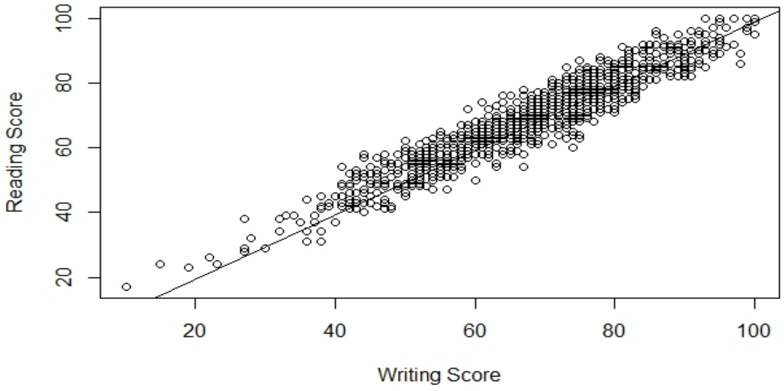


Figure 6: Scatter plot of Reading Scores against Writing Scores with regression line

Figure 6 above shows a regression model called simple linear regression which involves only one independent variable. The regression model is positive linear model which have a straight-line relationship. From the regression analysis, we can predict that when writing score increases, the reading score increases.

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

$\hat{y}$: Estimated (or predicted) y value
b_0: Estimate of regression intercept
b_1: Estimate of the regression slope
x : Independent variable

By using RStudio, we can form a regression model equation, $\hat{y} = -0.6676 + 0.9935x$**.** Based on the equation, we can know that the estimated changes of the average reading score will increase by 0.9935 when the writing score increases. -0.6676 is the estimated reading score when the writing score is 0.

From RStudio, we also can know that, the value of coefficient of determination, $R^2 = 0.9113$. This means that 91.13% of the variation in Reading Score (dependent variable) is interpreted by Writing Score (independent variable).

3.1.4 Chi Square Test

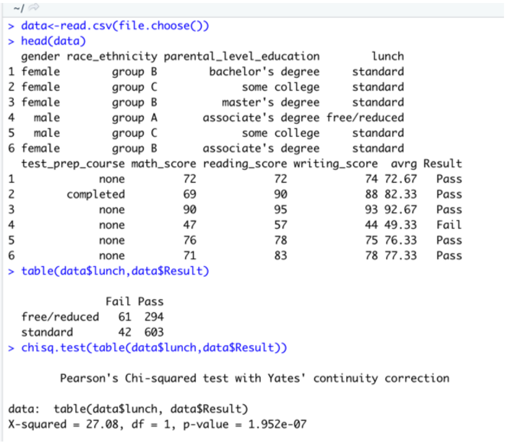
By using the Chi Square test, we are now investigating the relationship between the lunch and the students' performance in examinations.

According to the students' score from the secondary data, the average score of three subjects is calculated and it is categorized into 2 categories : pass and fail.

Let H_o = null hypothesis and H_1 = alternative hypothesis. The hypothesis are shown below:

H_o: There is no relationship between the lunch and the students' performance in examinations.

H_1: There is a relationship between the lunch and the students' performance in examinations.



```
~/
> data<-read.csv(file.choose())
> head(data)
  gender race_ethnicity parental_level_education        lunch
1 female        group B        bachelor's degree     standard
2 female        group C             some college     standard
3 female        group B          master's degree     standard
4   male        group A       associate's degree free/reduced
5   male        group C             some college     standard
6 female        group B       associate's degree     standard
  test_prep_course math_score reading_score writing_score  avrg Result
1             none         72            72            74 72.67   Pass
2        completed         69            90            88 82.33   Pass
3             none         90            95            93 92.67   Pass
4             none         47            57            44 49.33   Fail
5             none         76            78            75 76.33   Pass
6             none         71            83            78 77.33   Pass
> table(data$lunch,data$Result)

               Fail Pass
  free/reduced   61  294
  standard       42  603
> chisq.test(table(data$lunch,data$Result))

	Pearson's Chi-squared test with Yates' continuity correction

data:  table(data$lunch, data$Result)
X-squared = 27.08, df = 1, p-value = 1.952e-07
```

Figure 7: Chi Square test of RStudio

By using RStudio, we can obtain:

	Fail	Pass
free/reduced	61	294
standard	42	603

X^2 = 27.08 , df = 1 , p-value = 0.0000001952

The hypothesis is tested using a 5% significance level. To reject the null hypothesis, X^2 value must be greater than 3.841, while p-value must be less than 0.05.

From the computed result, we can conclude that:

P-value = 0.0000001952 < 0.05 and X^2 = 27.08 > 3.841, therefore, Reject $H_{0.}$ There is sufficient evidence to conclude that there is a relationship between the lunch and the students' performance in examinations.

Conclusion

In this project, we have learned how to use RStudio to conduct hypothesis test, correlation test, regression test and Chi Square test. We also learn to construct a plotted graph using RStudio. From the result shown in RStudio, we can make conclusion on the studied data and determine the relationship of the variables.

From the analysis, we can conclude that the population mean of average score is greater than 50.00. This means that the overall performance of students in this examination is higher than the passing line. In correlation test, there is a relationship between the mathematics scores and the writing scores among the students. While in regression test, there is a relationship between writing scores and reading scores of the students. The average reading score will increase by 0.9935 when the writing score increases. 91.13% of the variation in Reading Score is interpreted by Writing Score. In the chi-squared test, we also know that there is a relationship between there is a relationship between the lunch and the students' performance in examinations.

Appendix

Raw dataset :
StudentsPerformance.csv