SECI2143 : PROBABILITY AND STATISTICAL DATA ANALYSIS

2021/2022 - SEMESTER 2

PROJECT 2 (10%)

No.	STUDENT NAME	MATRIC NUMBER
1	LIM JIE HAN	A21EC3013
2	JELIZA JUSTINE A/P SEBASTIN	A21EC0034
3	THUVAARITHA A/P SIVARAJAH	A21EC0137

TABLE OF CONTENTS

CONTENT	PAGE
1.0 Introduction	3
2.0 Objective	3
3.0 Methodology	3
4.0 Descriptive Statistics	4
5.0 Inferential statistics	5
6.0 Statistical Data Analysis	6,7
7.0 Interpretation of Results	8-17
7.1 Hypothesis Testing-One Sample Test	8,9
7.2 Correlation	10,11,12
7.3 Regression	13,14,15
7.4 Chi-Square Test	16,17
8.0 Conclusion and Discussion	18
9.0 Work Coordination	19
10.0 Reflection	20,21
11.0 References	22
12.0 Appendices	23

1.0 Introduction

An examination (exam) is a test. Many things may be examined, but the word is most often used for an assessment of a person. It measures a test-taker's knowledge, skill, aptitude, physical fitness, or ability or standing in some other topic. It is a set of questions designed to measure those things. With every pupil being so individual, exams are also a great way for teachers to find out more about the students themselves. The test environment comes with added stress, which allows teachers to work out how their students argue and how they think individually by their work, which is a great attribute for them to keep in mind for future class activities.

Academic performance is among the several components of academic success. Many factors, including socioeconomic status, student temperature and motivation, peer, and parental support influence academic performance. This particular dataset contains data from a sample of 40 students. This study is aimed to investigate the average score scored in tests among students and determine the factors that affect their marks.

2.0 Objective

The objective of this study is to investigate the relationship of average scores obtained in a test. This test is conducted by performing and finding inferential statistics such as one sample hypothesis testing, correlation test, regression test and chi-square test. Certain variables are chosen for each specific test.

3.0 Methodology

This dataset was retrieved from an online open source website called Kaggle which provides and collects data. The data used in these tests are secondary data. We have decided to carry out 4 different tests on variables from the data source. From the data, we have selected 40 sample datas at random and have carried out 1 sample hypothesis testing, correlation test, regression test and chi-square test. These statistical tests are done to find out the relationship between the variables. The calculations and plots are plotted using RStudio.

4.0 Descriptive Statistics

The dataset that we have chosen for our project is the student's performance in the examination dataset. This dataset contains data from a sample of 40 students. The variable in the dataset is shown in Table 1 and the summary of the selected variables and the test with their respective description is shown in Table 2.

Raw Data

Variable	Data Type
Gender	Nominal
Race/ethnicity	Nominal
Parental level of education	Ordinal
Lunch	Nominal
Test preparation course	Nominal
Math score	Ratio
Reading score	Ratio
Writing score	Ratio

Table 1

5.0 Inferential Statistics

Selected variable(s)	Test	Description
Math score	Hypothesis testing (one sample t-test)	The variable is used to test if the average score is not equal to 73.66 at 95% confidence interval.
Reading score, Writing score	Correlation	The variables are used to test if the true correlation is not equal to 0 at 95% confidence interval.
Reading score, Writing score	Regression	The variable writing score is used as an independent variable and the variable reading score is used as a dependent variable. The variables are used to show the linear relationship between reading and writing scores.
Gender, Test preparation course	Chi square test	The variables used to determine the relationship between gender and test preparedness.

Table 2

6.0 Statistical Data Analysis

```
library(readxl)
> score <- read_excel("C:/Users/User/Desktop/score.xlsx")
> View(score)
> attach(score)
> # one sample t-test
  > t.test(`math score`)

One Sample t-test

data:  math score
t = 41.598, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  70.0212 77.1788
sample estimates:
  mean of x
73.6

> # correlation
  > cor.test(`reading score`,`writing score`)

Pearson's product-moment correlation

data:  reading score and writing score
t = 16.108, df = 38, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8778187 0.9647697
sample estimates:
      cor
0.9339426

> plot(`reading score`,`writing score`)
> # regression
> reg=lm(`reading score`~`writing score`)
> summary(reg)
```

```
Call:
lm(formula = `reading score` ~ `writing score`)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1843  -2.2191   0.3023   2.4414   8.8852

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.37447    4.53186   0.965    0.341
`writing score`  0.96524    0.05992  16.108   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.288 on 38 degrees of freedom
Multiple R-squared:  0.8722,  Adjusted R-squared:  0.8689
F-statistic: 259.5 on 1 and 38 DF,  p-value: < 2.2e-16

> # Chi square test
> table(gender)
gender
female   male
    19     21
> table(`test preparation course`)
test preparation course
completed      none
       23        17
> table(gender,`test preparation course`)
        test preparation course
gender   completed none
  female        10    9
  male          13    8
> chisq.test(gender,`test preparation course`)

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

data:  gender and test preparation course
X-squared = 0.074098, df = 1, p-value = 0.7855
```

7.0 Interpretation of Results

7.1 Hypothesis Testing - One Sample Test

Previously, Sunway university conducted a survey on the average score scored by students in the Math examination in SPM. From the survey they found that the average mean score scored by students is 73.66 out of 100 marks. Therefore, the null hypothesis, H0 and alternative hypothesis, H1, is as follows:

$$H_o: \mu = 73.66$$

$$H_1: \mu \neq 73.66$$

Where μ indicates the average score scored by students in the Math examination.

A random sample of 40 students were randomly collected from the secondary data source. It is calculated that the sample data has a mean of 73.6 marks of Math score. We can calculate the standard deviation by using the following formula:

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})2}{n-1}}$$

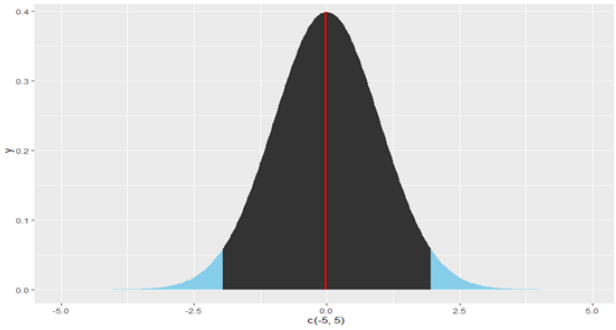
From the calculation, we can get the value of standard deviation which is 11.1902.

The Math score variable is used to test the average score is not equal to 73.66 at 95% confidence interval. The critical value of 95% confidence interval is 1.96 and -1.96. So, we can calculate the z-value of mean by using the following formula:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

From the calculation we can have the value of z is -0.03391.

$\bar{x}$	73.6
μ	73.66
s	11.1902
z-value	-0.0339
Critical value	-1.96 +1.96



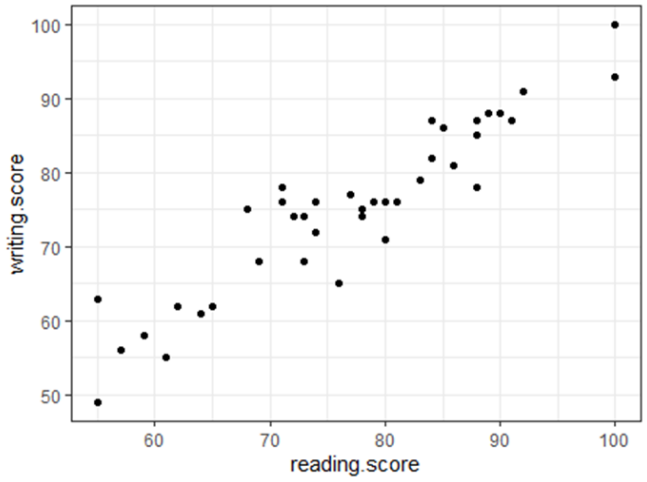
The z-value which is 0.03391 falls in between the critical region because it is less than 1.96 but more than -1.96. From the calculation and graph, we fail to reject the null hypothesis, H0 because we have insufficient evidence to support the claim that the average score scored by students in Math examination is not equal to 73.66 marks.

7.2 Correlation

The correlation test is taken to see the relationship between reading and writing score from the data collected. Both the variables are in ratio form so the coefficient correlation is calculated using Pearson's technique. It is referred to as a bivariate variable. The variable reading score is used as an independent variable and the variable writing score is used as a dependent variable. The X variable is reading score and the Y variable is writing score. The coefficient correlation is calculated using the following formula:

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

From the calculation, we can get the result 0.9339426. We can make a statement that the relationship between both these variables are strong and positive. The test result shows that when writing score increases the reading score also increases. The scattered plot shows a positive slope.



∑x	3062
∑y	2991
∑x2	9375844
∑y2	8946081
∑xy	233904
r	0.9339426

Significance test for correlation

Both variables are used to test whether there is any evidence of linear relationship between them at 95% confidence interval. The null hypothesis, H0 is assumed that the reading score has no correlation.

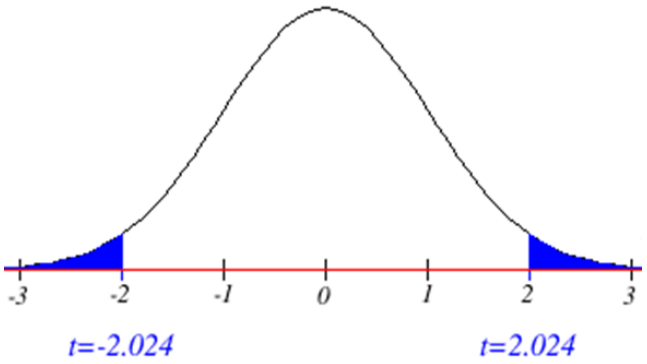
H0: $\rho = 0$; p = population correlation coefficient

H1: $\rho \neq 0$

Using the formula, we can calculate the test statistics where the value of t is 16.108 with the degree of freedom 38. The critical values are -2.024 and 2.024

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

```
Pearson's product-moment correlation

data:  reading score and writing score

t = 16.108, df = 38, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.8778187 0.9647697

sample estimates:

      cor

0.9339426
```



The graph shows the test statistics value falls in the critical region. We will reject the null hypothesis at 95% confidence interval since the test statistics value t is 16.108 is greater than the critical value which is 2.024. Because there is sufficient evidence to claim that there is a linear relationship between the dependent and independent variables at 95% confidence interval and we also have sufficient evidence to claim that the population correlation coefficient, p is not equal to zero.

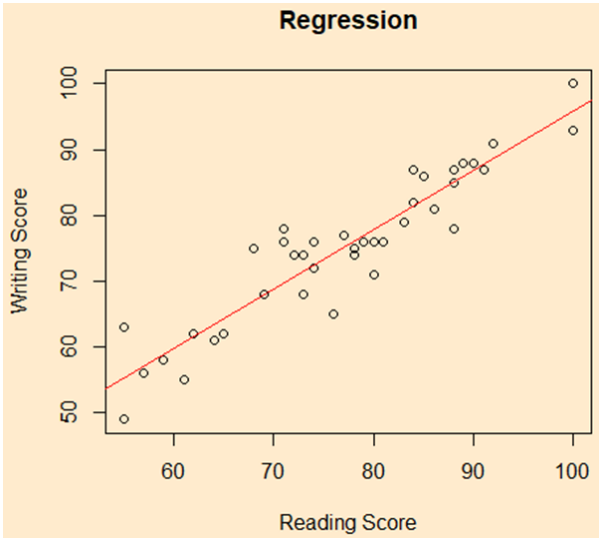
7.3 Regression

The variable writing score is used as an independent variable and the variable reading score is used as a dependent variable. The variables are used to show the linear relationship between reading and writing scores.

The mathematical equation for this regression is as follows:

$Y = \beta 0 + \beta 1X + \varepsilon$

Y is dependent variable,
β0 is the population intercept y,
β1 is the population slope coefficient
X is independent variable,
ε is a random error component.



The graph of linear regression shows the relationship between reading and writing scores.

```
Call:
lm(formula = `reading score` ~ `writing score`)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1843  -2.2191   0.3023   2.4414   8.8852

Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.37447    4.53186   0.965    0.341
`writing score`  0.96524    0.05992  16.108   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.288 on 38 degrees of freedom
Multiple R-squared:  0.8722,  Adjusted R-squared:  0.8689
F-statistic: 259.5 on 1 and 38 DF,  p-value: < 2.2e-16
```

From the calculation above, the estimate of regression intercept is 4.37447 and the estimated regression slope is 0.96524. The estimated regression model can be calculated by $\hat{y} = 4.37447 + 0.96524x$.

From the calculation we also obtain the residual standard error 4.288 on 38 degrees of freedom. The value of R2 (R square) is 0.8722.

The test statistical is calculated by the following formula:

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

The b1 is 0.96524 from the previous calculation. The Sb1 is the estimator of the standard error of the slope.

The following formula is used to calculate the Sb1:

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

The following formula is used to calculate the Se value:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \text{Sample standard error of the estimate}$$

The following formula is used to calculate the value of SSE:

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

Se = 4.288 Sb1 = 0.05992
Test statistic, T = 16.108
+T (38, 0.025) = 2.0244
-T (38, 0.025) = -2.0244

The test statistic (16.108) is > than the critical region (38, 0.025) = 2.0244. So, we reject the null hypothesis. There is sufficient evidence to claim that there is a linear relationship between reading and writing scores at 95% confidence interval.

7.4 Chi Square Test

The chi square test is used to determine the relationship between gender and test preparedness.

H0: There is no relationship between gender and test preparedness and both variables are independent of each other.

H1: There is a relationship between gender and test preparedness and both variables are dependent on each other.

```
> table(gender)
gender
female   male
    19     21
> table(`test preparation course`)
test preparation course
completed      none
       23        17
> table(gender,`test preparation course`)
        test preparation course
gender   completed none
  female        10    9
  male          13    8
> chisq.test(gender,`test preparation course`)

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

data:  gender and test preparation course
X-squared = 0.074098, df = 1, p-value = 0.7855
```

The observed and expected values are shown in the table below.

Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}
1,1	10	(19 x 23)/40 =10.925
1,2	9	(19 x 17)/40 =8.075
2,1	13	(21 x 23)/40 =12.075
2,2	8	(21 x 17)/40 =8.925

The test statistic, chi square value obtained is 0.074098. The degree of freedom is 1 and the p-value is 0.7855.

The test statistic value does not fall in the critical region. X2 = 0.0741 is < than X2 (1,0.05) = 3.841. We fail to reject the null hypothesis because there is sufficient evidence to prove the relationship between gender and test preparedness at a 95% confidence interval.

8.0 Conclusion and Discussion

Based on the one sample test, we fail to reject the null hypothesis. There is insufficient evidence to support the claim that the average score scored by students in the Math examination is not equal to 73.66 marks.

Next, from the correlation test, it is found that there is a strong positive relationship between the average writing score and average reading score with a correlation coefficient (r) of 0.9339426. We will reject the null hypothesis at 95% confidence interval since the test statistic value t = 16.108 is greater than the critical value t = 2.024. Due to the fact that there is sufficient evidence to claim that there is a linear relationship between the dependent and independent variables at 95% confidence interval and we also have sufficient evidence to claim that the population correlation coefficient, p is not equal to 0.

The estimated regression model is then produced from which we obtain $\hat{y} = 4.37447 + 0.96524x$ and this regression model is helpful in predicting the average reading score based on average writing score. The test statistic (16.108) is > than the critical region (38, 0.025) = 2.0244. So, we reject the null hypothesis. There is sufficient evidence to claim that there is a linear relationship between reading and writing scores at 95% confidence interval.

As for the chi square test, the test statistic chi square value obtained is 0.074098. The degree of freedom is 1 and the p-value is 0.7855. The test statistic value does not fall in the critical region. $X^2 = 0.0741$ is < than X^2 (1,0.05) = 3.841. Therefore, we fail to reject the null hypothesis because there is sufficient evidence to prove the relationship between gender and test preparedness at a 95% confidence interval.

9.0 Work Coordination

NAME	MATRIC NUMBER	TASK
Lim Jie Han	A21EC3013	Calculation and plotting using Rstudio
Jeliza Justine A/P Sebastin	A21EC0034	Report : ● 4.0 ● 5.0 ● 6.0 ● 7.0
Thuvaaritha Sivarajah	A21EC0137	Report : ● 1.0 ● 2.0 ● 3.0 ● 8.0 ● 9.0

10.0 Reflection

Name	Reflection
Jeliza Justine A/P Sebastin A21EC0034	From this project I have learned many important things and gained knowledge from conducting such a project. Mainly, I have learned how to conduct an inference statistical analysis on a selected dataset. We have used the test hypothesis testing, correlation, regression, and chi square test of independence. Here, I thank my lecturer Dr Azurah for giving us an opportunity to carry out this project. In this project, I was able to implement all that I studied. My group and I faced many challenges while doing this project. First, we had difficulties in understanding the project and we had many questions on how to conduct this project. Our lecturer helped us to conduct this project by completing the syllabus first so that we can have a better understanding on how to conduct this project. She also provided a guideline on what we should do and where we should start. It was very helpful for me and my group. We also faced a hard time in selecting a proper dataset. We were not able to use the dataset that we chose first because it contains negative values. So, we took time to find a suitable dataset. Finally, we managed to find a dataset which was retried from Kaggle, an online open source data source. Moreover, I also learned more about R programming. At first, I felt overwhelmed with the tasks provided but I managed to tackle each task with the help of my group members. I browsed the internet to have some ideas and that's how I completed my tasks. I also thank my team members for their contribution in completing this project. We divided our tasks so that we don't feel overwhelmed. We always discuss our project if we have any doubt so that we don't miss out on anything. Teamworking plays an important role and that was the main strength in completing this project.

Name	Reflection
Lim Jie Han A21EC3013	In this project, I learn how to do the inferential statistics. I did one sample testing, correlation, regression and chi square test of independence. I am grateful to have such a great opportunity to do the analysis for the real life issue. It is totally different from the other assignment. I would like to thank my lecturer Dr Azurah for her guidance and assistance. From her lecture, I gained a lot of knowledge about the inferential statistics and learnt the methods how to apply them in this project. She also explained about this project to us and her briefing is very helpful for us to finish this project. This project is very interesting because we can know more about the issue that we have studied after we analyze it. In my opinion, from this project and this course, I can prepare myself for the analysis work in the future. Besides that, I learnt more about R programming and I think that it is very helpful for statistical analysis work. I also thank my team members for their contribution in the progress of finishing this project. Teamwork is very important for conducting this project.

Name	Reflection
Thuvaaritha A/P Sivarajah A21EC0137	From this project, I understood the significance of the four tests even better. The four tests include one sample test, correlation test, regression test and the chi square test. The dataset we chose is a very relatable one and definitely a topic we're familiar with hence, making it much more fun to work with and analyze. This project also helped with my data analyzing skills a little more.In all honesty, choosing a proper dataset was probably the hardest task because we could never seem to get a perfect one. However, after much looking, we found a good one that we could work with. I'd like to take this opportunity to thank Dr.Azurah for all her guidance throughout this project especially at the beginning when we had no clue and did not know where or how to start. I'd also like to thank my teammates for their teamwork and never ending support. It was extremely easy working with them and I'm grateful to have worked together.

11.0 References

Dataset : https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Other reference :
https://www.researchgate.net/publication/288835170_SECONDARY_SCHOOL_MATHEMATICS_GENDER_AND_MUFY_MATH_PERFORMANCE_A_SUNWAY_CAMPUS_CASE_STUDY

List of e-portfolios :

NAME	MATRIC NUMBER	E-PORTFOLIO LINK
Lim Jie Han	A21EC3013	https://eportfolio.utm.my/user/lim-jie-han
Jeliza Justine A/P Sebastin	A21EC0034	https://eportfolio.utm.my/user/jeliza-justine-a-p-sebastin
Thuvaaritha Sivarajah	A21EC0137	https://eportfolio.utm.my/user/thuvaaritha-eportfolio

Link of presentation video :
https://www.canva.com/design/DAFEyOZwIMk/GlU55M6allMEtkaqIKYb7Q/edit?utm_content=DAFEyOZwIMk&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

12.0 Appendices