UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 - 02

PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

DR. CHAN WENG HOWE

GROUP 11 (SURVEY CORPS)

NAME	MATRIC NO.
ADAM WAFII BIN AZUAR	A20EC0003
AHMAD MUHAIMIN BIN AHMAD HAMBALI	A20EC0006
FARAH IRDINA BINTI AHMAD BAHARUDIN	A20EC0035
MUHAMMAD NAQUIB BIN ZAKARIA	A20BE0161

1. INTRODUCTION

Every social work education programme evaluates a student's performance to defined standards and criteria and the learning objectives and techniques described in the student's learning contract. Evaluations of student performance have two types which are informal and formal. Informal evaluation is dependent on the instructor, whereas formal evaluation is based on student examination performance.Ongoing monitoring is very important to ensure the student progress, identify strengths, and to identify areas of performance that may need special attention. Most of the countries in the world have implemented this method in their formal learning to produce quality students. By doing this method, the student can identify their weaknesses and strengths during their studies. This method will be beneficial when the student needs to choose their domain in higher-level education.This study investigates the relationship between writing performance and reading performance time from data provided and determines whether the reading performance will affect the writing performance.

2. DATA SET

The data in this study is obtained from the dataset “Students Performance”. This dataset contains the data of the scores of Mathematics, Reading and Writing. The data also provides the number of students, gender, race and ethnicity, parental level of education and type of lunch.The 500 data which are respective to the 500 students are selected from the database and used for hypothesis testing to determine whether there is enough evidence to support the null hypothesis. The sample is normally distributed and plotted with RStudio. The datasets contain five qualitative datas which are gender, race or ethnicity, parental level of education, lunch and test preparation course. It also contains three quantitative datas which are math score, reading score and writing score.

Variable	Data Type	Level of Measurement
Gender	Qualitative	Nominal
Race/Ethnicity	Qualitative	Nominal
Parental level of education	Qualitative	Nominal
Type of lunch	Qualitative	Nominal
Math score	Quantitative	Ratio
Reading score	Quantitative	Ratio
Writing score	Quantitative	Ratio

3. DISCUSSION AND RESULT

A. Hypothesis

Based on previous study, the survey found that the average math score for the 500 students is more than 65.

Hence, the null hypothesis, H_0 and alternative hypothesis, H_1 is:

H_0: $\mu = 65$

H_1: $\mu > 65$

Where μ is the mean of the average math score for the students.

A random sample of 500 math scores produced from the students having a mean 65.714. The standard deviation can be evaluate using this formula:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

The standard deviation is 15.0404

A 5% level of confidence is used to test the claim of this study that the average math score of the students is more than 65. The critical value of 5% level of confidence is -1.645. the z-value of mean can be calculated by using below formula:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

And z is equal to 1.0615

$\overline{X}$	μ	s	z-value	Critical Value
65.714	65	15.0404	1.0615	-1.645

Since the z-value (1.0615) > critical value (1.645), we reject the null hypothesis. Hence, there is sufficient evidence to support the claim that the average math score of the students is more than 65.

B. Correlation

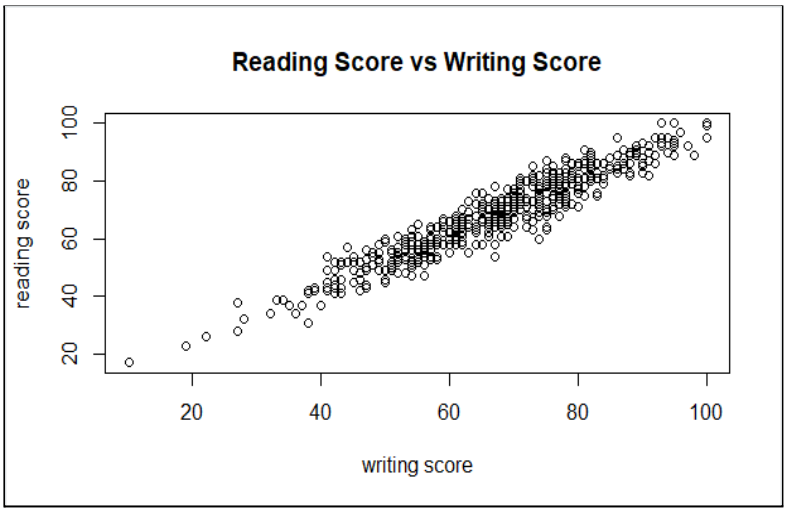


Fig 1: Scatter plot of reading score against writing score

A scatter plot or scatter diagram shows the relationship between two variables: reading score and writing score. The data pattern will indicate the relationship type, either positive or negative, with no relationship and curvilinear relationship. In this correlation test, we can measure the strength of the association between reading scores and writing scores with a sample of 500 students. We did the correlation coefficient, r, the calculation by using Rstudio. The provided data was in ratio level. Therefore, we calculate the correlation coefficient by using Pearson's technique. Based on the scatter plot produced, we can see that the relation between writing score and reading score is a positive relationship. From the calculation, the value of r is 0.9530237 indicates that variables x and y have a strong relation. However, we cannot conclude that the better the student in the writing test, the better the student in the reading test, but it is enough evidence to show that both reading and writing score correlate.

```
        Pearson's product-moment correlation

data:  StudentsPerformance$`reading score` and
 StudentsPerformance$`writing score`
t = 70.214, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not
 equal to 0
95 percent confidence interval:
 0.9442448 0.9604485
sample estimates:
      cor
0.9530237
```

Fig 3: Calculation of correlation coefficient in R

B.1. Significance Test for Correlation

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

H_1 : $\rho \neq 0$

In this test, the test statistics, where sample size n = 500 and r equals 0.9530237 as calculated above. Value of t can be obtained by using the R which is 70.214.

Test Statistic, t	α	Degree of Freedom, v	Critical Value
70.214	0.05	498	1.96473898 or -1.96473898

Since the test statistic, t > critical value (70.214>1.965), we reject the null hypothesis at α = 0.05. There is sufficient evidence to claim that there is a linear relationship between the writing score and reading score at 5% level of significance.

C. Linear Regression

Linear regression is one of the most commonly used predictive modelling techniques which describes the relationship between the independent variable and dependent variable as a straight line. Only one independent variable, x. Thus, the aim of linear regression is to predict the y-value based on the x-value. Changes in y are assumed to be caused by changes in x.

This mathematical equation can be generalized as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y is dependent variable,

β_0 is population y intercept,

β_1 is population slope coefficient,

x is independent variable, and

ε is a random error term or residual.

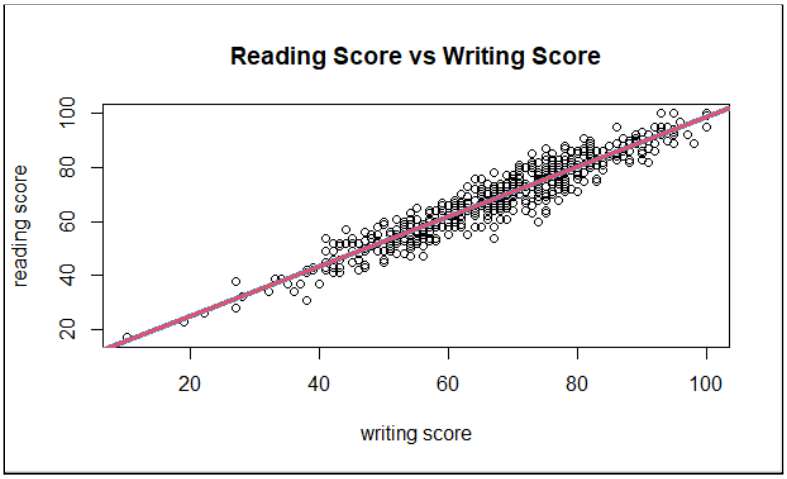


Fig 4: Scatter plot of reading score against writing score

Referring to figure 4, regression shows the result of the relationship between writing score and reading score. It can be seen that the graph has a positive linear relationship between reading score and writing score as the points are closely scattered along the straight line.

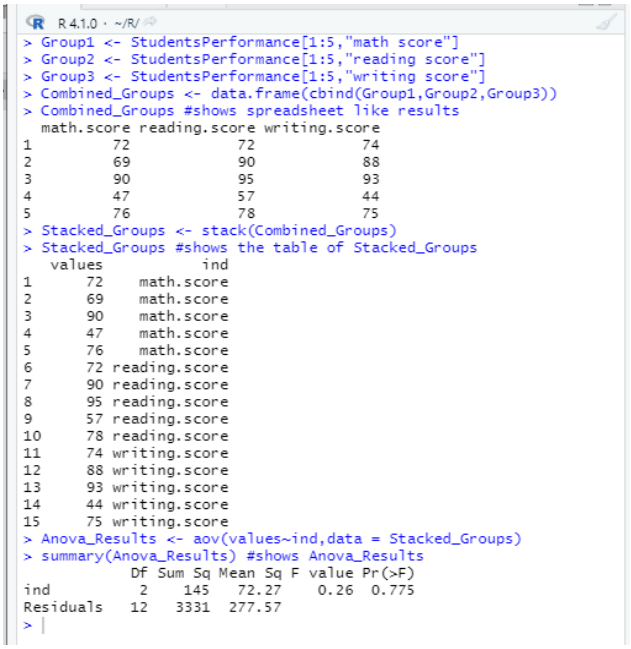
```
Call:
lm(formula = `reading score` ~ `writing score`)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6716  -3.0666   0.1309   3.1233  11.4074

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.51718    0.90497   7.202 2.21e-12 ***
`writing score`  0.92101    0.01312  70.214  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.472 on 498 degrees of freedom
Multiple R-squared:  0.9083,    Adjusted R-squared:  0.9081
F-statistic:  4930 on 1 and 498 DF,  p-value: < 2.2e-16
```

Fig 5: Summary of regression test

The estimated regression model is then calculated using RStudio, where we obtain $\hat{y}$ = 6.51718 + 0.92101x. Based on the result, every 1 mark of writing score will cause the reading score to increase by 0.92101. This regression test is said to be statistically significant because the p-value is very close to zero which is smaller than the significant level, 0.05.

Moreover, the coefficient of determination, R^2 is calculated to see the portion of the total variation in the dependent variable, reading score is explained by the variation in the writing score. R-square (R^2) is then calculated by using RStudio in which we obtained R^2 is 0.9083 which is 90.83% of the variation in reading score is explained by the variation in the writing score.

D. ANOVA



```
R 4.1.0 · ~/R/
> Group1 <- StudentsPerformance[1:5,"math score"]
> Group2 <- StudentsPerformance[1:5,"reading score"]
> Group3 <- StudentsPerformance[1:5,"writing score"]
> Combined_Groups <- data.frame(cbind(Group1,Group2,Group3))
> Combined_Groups #shows spreadsheet like results
  math.score reading.score writing.score
1         72            72            74
2         69            90            88
3         90            95            93
4         47            57            44
5         76            78            75
> Stacked_Groups <- stack(Combined_Groups)
> Stacked_Groups #shows the table of Stacked_Groups
   values         ind
1      72  math.score
2      69  math.score
3      90  math.score
4      47  math.score
5      76  math.score
6      72 reading.score
7      90 reading.score
8      95 reading.score
9      57 reading.score
10     78 reading.score
11     74 writing.score
12     88 writing.score
13     93 writing.score
14     44 writing.score
15     75 writing.score
> Anova_Results <- aov(values~ind,data = Stacked_Groups)
> summary(Anova_Results) #shows Anova_Results
            Df Sum Sq Mean Sq F value Pr(>F)
ind          2    145   72.27    0.26  0.775
Residuals   12   3331  277.57
>
```

Figure 6 ANOVA test by RStudio

ANOVA method of testing is used to analyze the significant differences between means. One-way ANOVA with an equal sample size method is used. Figure shows the list data of Group1, Group2, and Group3 that we extract from the dataset "StudentPerformance". In that case, the dataset "StudentPerformance" is distributed into 3 different categories. Group1 stands for "math score" while Group2 stands for "reading score" and Group3 stands for "writing score". Significance level of 0.05 is used to test the null hypothesis that the 5 students have the same mean on 3 different academic scores which is math score, reading score, and writing score. The null hypothesis and alternative hypothesis as below:

H0 : $\mu 1 = \mu 2 = \mu 3$

H1: at least one mean is different.

Table of mean and standard deviation of Academic scores on 5 students.

Math score	Reading score	Writing score
$\bar{x} = 70.8$	$\bar{x} = 78.4$	$\bar{x} = 74.8$
n = 5	n = 5	n = 5

S = 15.547	S = 15.077	S = 17.882

Test statistic, F is calculated using RStudio and the formula is as below:

F= (variance between sample (nS 2/x))/(variance within sample (s_p^2))

We get the variance between samples, nS 2/x = 72.27 while variance within the sample, s_p^2 = 277.57. Then test statistic, F = 0.26. The numerator, k-1 = 3-1 = 2, while the denominator, k(n-1) = 3(5-1) = 12, then F-distribution table is F (2,12) = 3.885.

Since F test statistic < F critical value (0.26 < 3.885), the test statistic does not fall within the critical region, therefore we fail to reject the null hypothesis. There is sufficient evidence to claim that 5 students have the same mean academic score which is math score, reading score and writing score.