



# UTM

UNIVERSITI TEKNOLOGI MALAYSIA

**SECI2143 SECTION 02: PROBABILITY & STATISTICAL DATA ANALYSIS**





2020/2021 – SEMESTER 2

TEAM 8: STATS BUDDIES

PROJECT 2

LINK TO VIDEO: <https://youtu.be/oMcgBOc6CR0>

ENGINEERING FACULTY  
SCHOOL OF COMPUTING

NAME	MATRIC NO.	PHONE NO.	PICTURE
ADRINA ASYIQIN BINTI MD ADHA	A20EC0174	011-13033218	
MADINA SURAYA BINTI ZHARIN	A20EC0203	012-3136850	
NAYLI NABIHAH BINTI JASNI	A20EC0105	019-2463571	
NUR SYAMALIA FAIQA BINTI MOHD KAMAL	A20EC0118	011-63360800	

## **Introduction**

Predicting a student's success is a major worry for higher education administrators and it is not an easy task. The recordings of 1000 students, as well as their arithmetic, reading, and writing scores, are included in this dataset. They also tell us about their gender, cultural background, parent's education, whether they received free/reduced or regular meals, and whether or not they took any test preparation classes before their exams. As a result, we are interested in delving further into these recordings in order to better understand the relationship between these various surroundings or events and the students' performance. In addition, we also want to gain further knowledge in analysing the student performances dataset since it relates closely to our situations as a student.

The purposes of the study are to review and analyze the students' performance in various conditions and using the implementation of mathematical methods such as mean, standard deviation, hypothesis test, correlation test, regression analysis, and so on in explaining the dataset. Thus, we can improve the environment and solution for students to study.

## **Dataset**

The data is obtained from one of the datasets provided by our lecturer which is 'Student Performance' as it is the most suitable topic to our current condition as a student. From the dataset, there are 1000 sample sizes with 8 categories.

The first category is gender (male/female), followed by the race or ethnicity, which is divided into 4 groups (Group A, B, C, D) and each of them might be categorized by races in Malaysia such as Malay, Chinese, Indian and others. Thirdly is parental education level either bachelor's degree, master's degree, associate's degree, college or high school. Next is how they obtained their daily meal, either standard or free/reduced and also their test preparation status, either none or completed. Last but not least, there are also 3 types of numerical data to be compared, which are mathematics score, reading score and writing score.

From all the data provided, we can do the hypothesis testing to determine whether each type of data could support each other. This can be done by doing statistical proof to support the null hypothesis. The graph and data obtained also are mainly being produced by RStudio as the sample size is large.

## Data analysis

### 1. Hypothesis testing

In statistics, hypothesis testing is a method of determining whether or not the results of a survey or experiment are relevant. It essentially determines whether the results are valid by calculating the chances that they happened by chance. In hypothesis testing, we analysed the sum of three different scores which are math, reading and writing examinations between male and female students. To determine which group did better on the exam, 518 female students and 482 male students were assessed. We assume that this is a test on mean variance unknown in which the variance is equal, then calculate the result. Mean for the average scores and the standard deviation for female students are 208.7 and 43.6 respectively, while the male students scored the mean average of 197.5 and the standard deviation of 41.1 in the three examinations.

$H_0$  : There is no difference in the mean between male and female

$H_1$  : The female sample has greater mean than the male sample

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$DF = n_1 + n_2 - 2$$
$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Using the formula provided, we got the values of  $S_p = 42.413$ ,  $DF = 998$  and  $t = 4.173$ . The value of critical values using t-table with 95% level of significance is 1.645. Moreover, the p-value is  $1.593e-05$ . To conclude, since  $4.173 > (1.646)$ , we decide to reject the null hypothesis. There is significant evidence that the female sample has greater mean of the math, reading and writing scores than the male sample. Hence, female students perform better in the examinations than male students as they achieve higher grades and scores more.

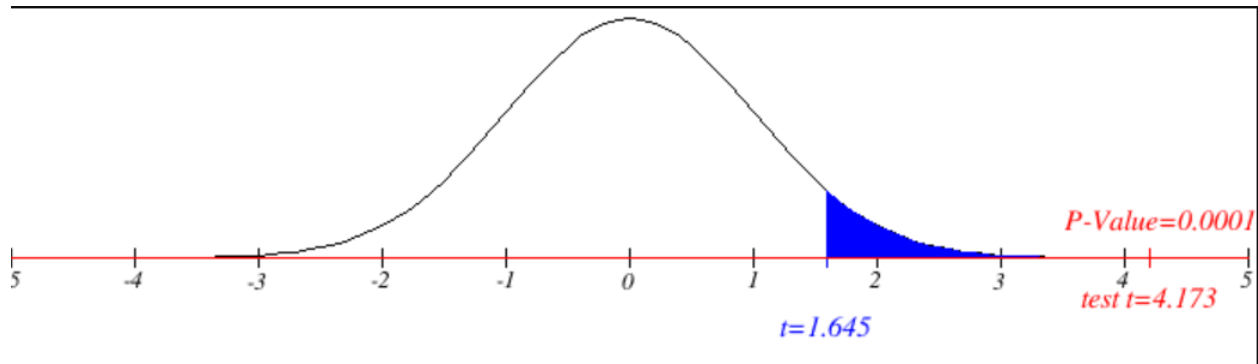


Figure 1: The t distribution graph

## 2. Correlation

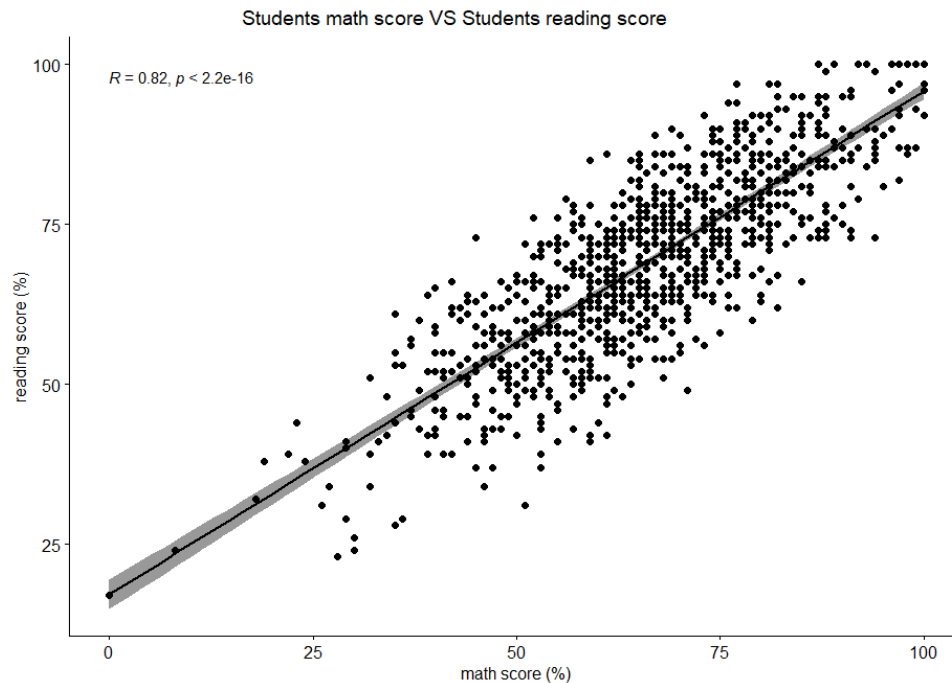


Figure 2: The correlation scatter plot reading

In this correlation test, we analysed the students performance using two variables which are math score and reading score presented in percentage. The relationship or strength of associations of the variables is tested with 1000 students as the sample size. The graph and the coefficient correlation,  $r$  is calculated using R Script to show a detailed relationship between both two. As both data are in ratio-type form, the coefficient correlation,  $r$  is calculated using Pearson's technique. The data are also comparable variables (bivariate data). By using Pearson's technique, formulae belows are being used.

Sample correlation coefficient:

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

where:

$r$  = Sample correlation coefficient

$n$  = Sample size

$x$  = Value of the independent variable

$y$  = Value of the dependent variable

The coefficient correlation,  $r = 0.8175797$  represents a strong positive correlation. As the number of math scores increase, the number of reading scores increase.

Significance Test for correlation

Now, we will test whether there is a linear relationship between math and reading score at the 0.05 level of significance. Assume that math score has no linear relationship as reading score by null hypothesis,  $H_0$ .

$H_0 : \rho = 0$ ,  $\rho$  = population correlation coefficient

$H_1 = \rho \neq 0$

Significance level,  $\alpha = 0.05$

Degree of freedom,  $df = 1000 - 2 = 998$

Sample size,  $n = 1000$

Coefficient correlation,  $r = 0.82$

Next, to find the test statistics, we use the formula:

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

Thus,  $t = 44.855$ . From the t-table, the critical value  $t_{0.05,998}$  is - 1.962344 and 1.962344. Since the test statistic,  $t > \text{critical value}$  ( $44.855 > 1.962344$ ), rejects null hypothesis,  $H_0$ . There is sufficient evidence of a linear relationship between math and reading score at the 5% level of significance.

### 3. Regression

Regression analysis often is conducted to predict one variable that lies on another variable. Regression always has at least two variables, one will act as a dependent variable while the other one will be an independent variable. If more variables exist, the dependent variable will always be only one and the other variables will be independent variables. There are many types of regression such as simple regression and multiple regression that have one and more independent variable(s) respectively. For this project, we will only do the simple linear regression analysis.

In this project, we use estimated regression model formula as shown below:

$$\hat{y}_i = b_0 + b_1 x \text{ where;}$$

$\hat{y}_i$  = estimated y value,

$b_0$  = estimated regression intercept,

$b_1$  = estimated regression slope,

$x$  = independent variable.

Regression analysis on 1000 students upon Reading Score (dependent) and Writing Score (independent)

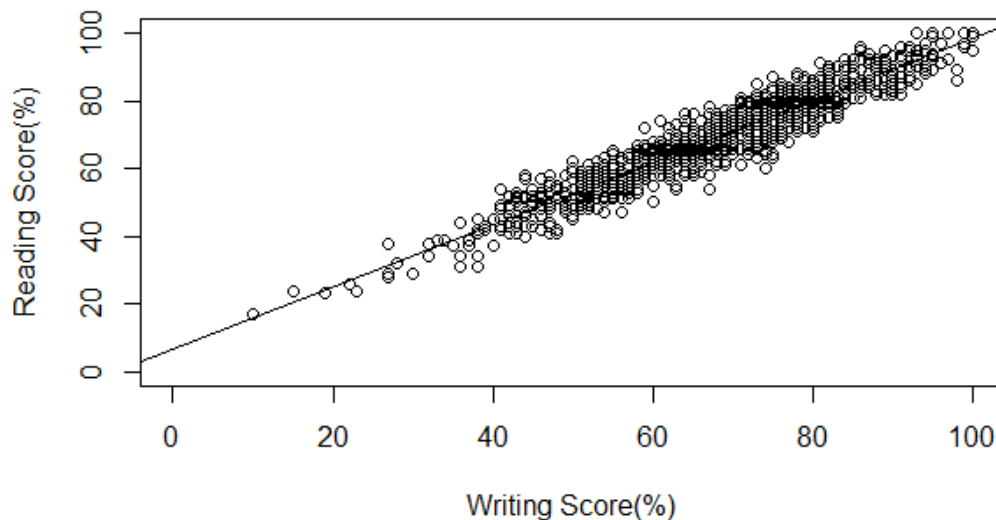


Figure 3: Scatter plot of reading score and writing score with regression line.

According to Figure 2 shown above, simple linear regression is done between the relationship on how reading score can be affected when writing score is manipulated. This means that the reading score is a dependent variable while the writing score is an independent variable. The figure also shows that it is a positive linear relationship.

The estimated regression model is calculated in RStudio, and it is found that  $\hat{y}_i = 6.7505 + 0.9172x$ . From the value stated above, the  $b_0$ , intersection coefficient, shows that no students get 0% on writing and reading scores. Therefore,  $b_0 = 6.7505$  indicates that 675 students from all 1000 students show that their reading score is not affected by their writing score. Next, the slope coefficient,  $b_1$ , is equal to 0.9172, which can be interpreted as the average reading score is increased as writing score also increases.

#### 4. Chi-square test of independence

Chi-Square test is performed to test if the hypothesis that math score is independent of parents education level fits the claim that math score is independent of parents educational level for student performances with a significance value of 0.05. The observed value are then compared to the corresponding expected values. The null hypothesis,  $H_0$  and alternative hypothesis,  $H_1$  is as follows:

$H_0$  : Math score is independent of parental level of education

$H_1$  : Math score is dependent of parental level of education

```
> colnames(parental_edu) ~ c("under_60", "60_to_80", "over_80",
> parental_edu <- as.table(parental_edu)
> parental_edu
      under_60 60_to_80 over_80
master's degree      9      28      22
bachelor's degree    20      63      35
associate's degree   58     109      55
some college        51     127      48
high school         73     101      22
some high school     63      84      32
> chisq <- chisq.test(parental_edu)
> chisq

Pearson's Chi-squared test

data:  parental_edu
X-squared = 45.477, df = 10, p-value = 1.784e-06

> chisq$observed
      under_60 60_to_80 over_80
master's degree      9      28      22
bachelor's degree    20      63      35
associate's degree   58     109      55
some college        51     127      48
high school         73     101      22
some high school     63      84      32
> chisq$expected
      under_60 60_to_80 over_80
master's degree  16.166  30.208  12.626
bachelor's degree 32.332  60.416  25.252
associate's degree 60.828 113.664  47.508
some college     61.924 115.712  48.364
high school     53.704 100.352  41.944
some high school 49.046  91.648  38.306
> qchisq(p = .95,df = 10)
[1] 18.30704
> |
```

Figure 4: Calculations obtained from RStudio for the Chi-Square test.

The observed values and expected values are shown in Fig. The test statistic, chi-square values then calculated using RStudio which equals 45.477 . The critical value can be found based on the Chi-Square Distribution table, with degrees of freedom equal to 10 and significance level of 0.05 . The p- value is found to be approximately  $1.784e-06$  which is equivalent to 0.000001784 .

The p-value is smaller than significance of 0.05, therefore we reject the null hypothesis. That is insufficient evidence that the student's math score is independent of parental level of education. In other words, student's math scores are dependent on parental level of education.

## **Conclusion**

This project has taught us to create and calculate hypotheses. Our group chose students' performance as our dataset because we are also students and we are interested in seeing the results and comparing our hypothesis and the calculated and proved hypothesis. Each of us had compared different objects and values to widen our view on the students' performances. The analysis process had us go in critical thinking mode, which requires deep thoughts and understanding on the topic. The project also helps us be more familiar with the usage of RStudio. The best results and findings from our project is when the parents educational background plays a role for the student's math scores. Never in a million years had we thought of these findings. After a while, we agreed with the results because parents with higher educational backgrounds might be more successful and have the choice to send their children to additional classes compared to parents who have lower educational backgrounds.