SECI2143

PROBABILITY AND STATISTICAL DATA ANALYSIS

SEMESTER 2 2020/2021

PROJECT 2

LECTURER:

DR SHARIN HAZLIN HUSPI

NO	NAME	MATRIC NO
1	ERICA DESIRAE MAURITIUS	A20EC0032
2	INDRADEVI A/P VIKNESHWARAN	A20EC0050
3	LUE GUO MING	A20EC0073
4	MOHD FIRDAUS BIN ZAMRI	A20EC0080

Table of Contents

Introduction

Nowadays, students' performance is one of the concern issues by parents, teachers or even the institution as the grades or the performance of a student will determine his or her future. Each student has their own strength in specific subjects. There are some students who excel in all subjects in school. However, there are also students who excel in one or few subjects. Moreover, according to many sayings, female students tend to have better performance compared to male students. Apart from that, we believe that there are also other factors influencing students' academic performance either long-term factors like family background and study environment or short-term factors like illness and emotion. Therefore, the aim of this project is to compare the academic performance between genders as well as to find out what factors are really affecting the students' academic performance and how strong are the relationship between those factors and students' academic performance. Besides that, we were interested in this topic because our curiosity hit us when we saw this topic as we would really like to know which gender does better in a specific academic subject and we are also wondering what type of student can obtain better academic results and the actual factor behind their academic success and achievement. In our expectation, we hope to see that most students can achieve a good academic performance with their own effort.

Methodology

For our research, our dataset named StudentsPerformance, is a secondary data that is collected from a Microsoft Excel provided by our Lecturer, Dr Sharin Hazlin Binti Huspi. The targeted population is students from educational institutions. We carry out the inferential statistics by using method hypothesis testing two samples, correlation, regression and ANOVA test.

Data Set

The dataset contains information of students and their subjects' results. From the dataset, there are continuous and discontinuous variables that we have identified. The continuous variables in the dataset are parental level of education, math score, reading score and writing score. Math score, reading score and writing score are from one category, which are under academic performance and they are the mark results scored by students. Parental level of education determines how high the educational level of students' parents had achieved. Meanwhile, gender, race/ethnicity, lunch and test preparation course are the discontinuous variables that we found in the dataset. In our assumption, lunch is some kind of scholarship that helps poor students, students with a free or reduced lunch will receive a discounted lunch, while students with standard lunch do not receive any lunch scholarship. In order to facilitate the research and get satisfactory results, some variables have been taken into consideration to carry out inferential statistics tests. The variables that to be considered are gender, test preparation, math score, reading score and writing score. In our research, variables such as gender and the average score of two tests (Writing and Reading) have been chosen to conduct a two-sample hypothesis testing to test whether both genders will have equal mean score of the three tests. Moreover, we also chose the variables the scores of three tests (Maths, Reading and Writing) and test preparations to perform a two-sample hypothesis testing to test whether students who have completed the test preparation have a higher mean score of the three tests

than the students who have not completed the test preparation. Besides that, we conduct a correlation test for the variables Math score, Reading score, Writing score to determine if there is a relationship between the scores. Variables maths score, reading score and writing score were chosen again to conduct a regression test as we assume reading score and writing score as a language score where it can predict the maths score. Finally, we conduct a one-way ANOVA test with an equal number of students to determine whether students will have balanced results or not for three different subjects. Table 1 shows the details in variables.

Variable	Suggested answer	Data type	Measurement level
Gender	male, female	Qualitative	Nominal
Race/ethnicity	Group A, group B, group C, group D, group E	Qualitative	Nominal
Parental level of education	Some high school, high school, some college, associate's degree, bachelor's degree, master's degree	Qualitative	Ordinal
Lunch	Standard, free/reduced	Qualitative	Nominal
Test preparation course	None, completed	Qualitative	Nominal
Math score	0-100	Quantitative	Ratio
Reading score	0-100	Quantitative	Ratio
Writing score	0-100	Quantitative	Ratio

Table 1

Data Analysis

Hypothesis Testing Two Sample

We have conducted hypothesis testing with two samples, in search of which gender does well in the language. The variables used for this test are gender and the language score. Meanwhile

the significance level is 0.05. In this hypothesis test, the variance is unknown, thus, we have proceeded by with an assumption that the unknown variance is unequal.

Name	Type	Value
t_test	list [10] (S3: htest)	List of length 10
statistic	double [1]	9.102857
parameter	double [1]	997.1068
p.value	double [1]	2.340251e-19
conf.int	double [2]	6.67 Inf
estimate	double [2]	72.5 64.4
null.value	double [1]	0
stderr	double [1]	0.8948321
alternative	character [1]	'greater'
method	character [1]	'Welch Two Sample t-test'
data.name	character [1]	's.language.female and s.language.male'

Figure 1 T test Summary

Based on Figure 1, we could have an overview of the T test conducted using RProgramming. The method used was the Welch Two Sample t-test.

Hypothesis statement:
H_0: female's mean language score is equal to male's mean language score
H_1: female's mean language score is greater than male's mean language score

Type of Hypothesis test: a right-tailed test

Test statistics:

T-statistics value, $t_0^* = 9.102857 \approx 9.103$

Degrees of freedom, $v = 997.1068 \approx 997$

Significance value used is $\alpha = 0.05$.

Critical value $t_{0.025,997} = -1.962$

Conclusion

Since $t_0^* > T_{0.025,997}$, whereby $9.103 > -1.962$. This proves that the test statistic value falls within the critical region. Thus, we reject the null hypothesis, H0. As a result, using a significance value of 0.05, there is sufficient evidence to conclude that the female's mean language score is greater than male's mean language score.

Correlation

We conduct a correlation test to find out whether there is a linear relationship between the score of reading and writing with a random sample of 1000 students. Hence, variables used in this test are reading score and writing score. We conduct this test with a significance level of 0.05.

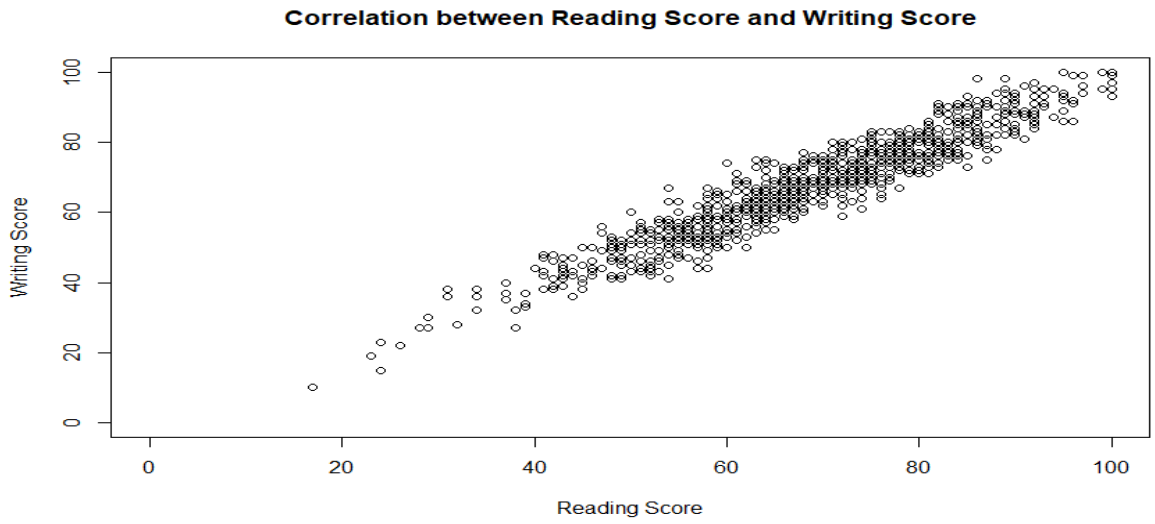

Figure 2 Scatter Plot of Writing Score and Reading Score

Based on Figure 2, the x-axis is the Reading Score and y-axis is the Writing Score. Moreover, from the scatter plot above, it can be seen that Writing Score increases as Reading Score increases. Hence, we can conclude that the scatter plot above shows a positive linear correlation.

By using Pearson's Product-Moment Correlation, the value of **r** is 0.9546, which indicates that a strong positive linear relationship exists between the two variables.

H_0: There is a linear correlation between Reading Score and Writing Score
H_1: There is no linear correlation between Reading Score and Writing Score

T-test value = 101.23

Degree of freedom = 998
Critical value = $t_{(0.025,998)}$ = 1.962368

Based on calculating t-test and critical value from R programming, t_0 = 101.23 and critical value $t_{0.025,998}$ = 1.962368. Thus, we can reject the null hypothesis since $t_0 > t0.025,998$. Therefore, by using a significance level of 0.05, there is sufficient evidence to conclude that a linear correlation exists between Reading Score and Writing Score.

Regression

For regression, we conduct a test to determine if language score can predict math score with a random sample of 1000 students. Hence, variables used in this test are language score and math score. We obtained the language score by calculating the mean score of reading score and writing score. We conduct this test with a significance level of 0.05. The dependent variable, y is the math score meanwhile the independent variable, x is the language score. The objective of this test is to test the existence of a linear relationship between y and x.

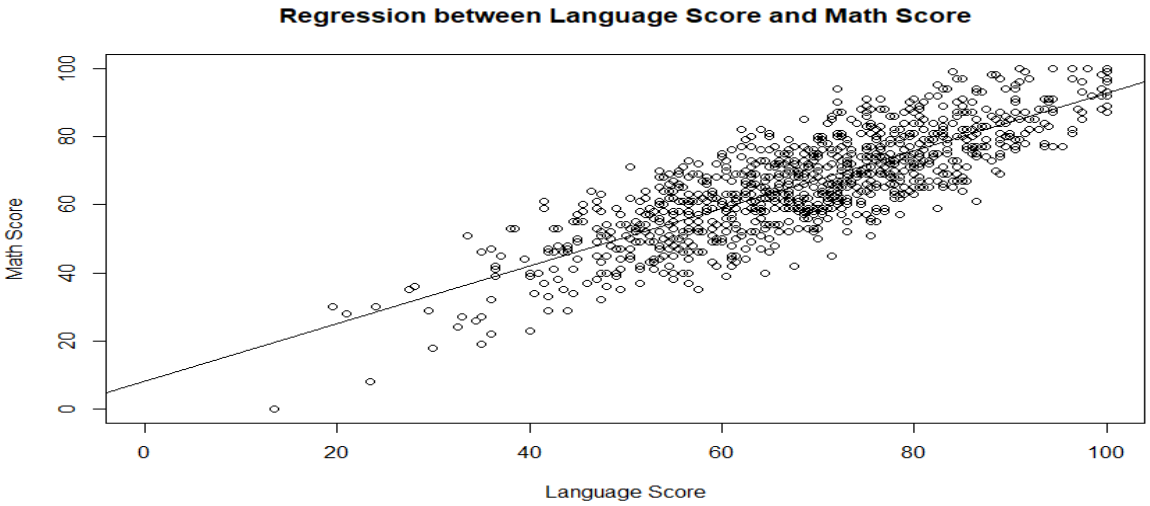

Figure 3 Scatter Plot of Language Score and Math Score

Based on Figure 3, the x-axis is Language Score and y-axis is Math Score. As we can see from the scatter plot, there exists a positive linear relationship among the variables. Through regression analysis, we can predict that the math score will increase as the language score increases. The equation obtained from this test is y = 8.2139 + 0.8435x.

The value of intersection coefficient, ß0, is the estimated value of y when x = 0. From the equation, the value of intersection coefficient is 8.2139 which indicates that if the language score is 0, then the value 8.2139 is the score for math. Meanwhile, for the value of slope coefficient, ß1, it measures the estimated change in the average value of y when x is changing one unit. The value 0.8435 shows that the average value of language score increased by 0.8435.

Coefficient of determination, R^2= 0.6713

H_0: ß1 = 0
H_1: ß1 ≠ 0

T-test value = 45.143
Degree of freedom = 998
Critical value = $t_{(0.025,998)}$ = 1.962368
Based on computing t-test and critical value from R programming, t_0 = 45.143 and critical value $t_{0.025,998}$ = 1.962368. Hence, we can reject the null hypothesis because $t_0 > t0.025,998$.

Thus, by using 0.05 significance level, we can conclude that there exists a positive relationship between dependent variable and independent variable.

ANOVA Test

Besides, we have conducted a One-Way Analysis of Variance with an equal sample size test to compare the scores among the three subjects (Maths, Reading and Writing). The mean was calculated from a sample size 1000 students data set for each subject.

Size of each sample, n = 1000
Number of populations being compared, k = 3
Numerator degrees of freedom = 2
Denominator degrees of freedom = 2997

H_0: mean of math score = mean of reading score = mean of writing score
H_1: at least one mean is different

The significance value used is $\alpha = 0.05$

The function aov which represents ANOVA test was utilized using the R Programming. A summary of the result is obtained. Using the result, box plots were plotted where the x-axis is the 3 subjects while the y-axis represents the score.

From the results, we obtained:
Sum of Squares = 4864
Mean Square = 2431.8
F value = 10.82
P-value of F-statistics = 0.0000207
Critical value of F with $\alpha = 0.05$ from F-distribution table = 2.996

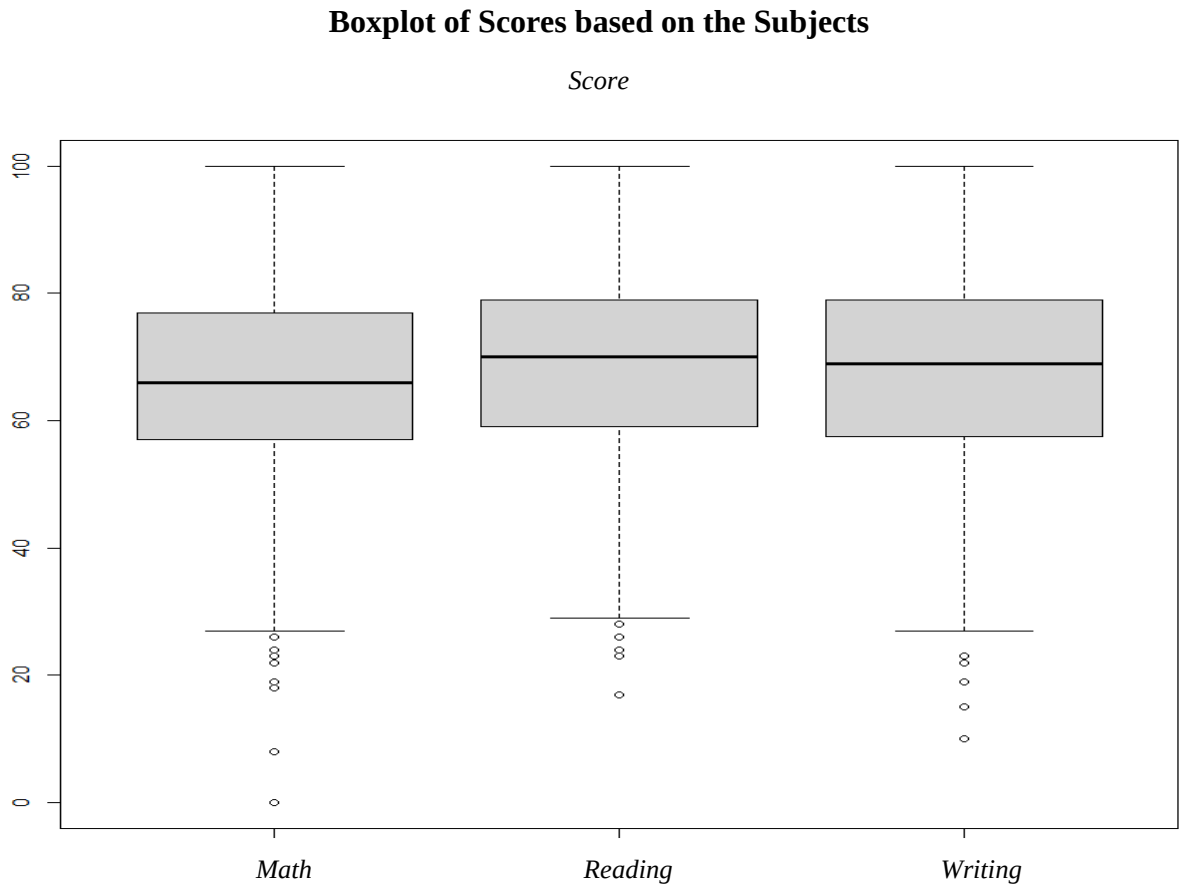

Figure 4 Box plots of the scores between the 3 subjects

Based on Figure 4, we can clearly see the difference of the mean between the 3 subjects. Besides that, math score has the most outliers compared to reading and writing score which affects the mean score.

Since the F test statistic > F critical value (10.82 > 2.996), we reject the null hypothesis, H_0. Thus, with a significance level of 0.005, there is sufficient evidence to claim that the mean scores for the subjects are different.

Conclusion

In this project, our group chose the dataset about student performance because it is related to us as we are students. Our group opine that we can have more ideas on possible analysis that relate to students' academic performance. During data pre-processing, our group brainstormed for lots of possible analysis and hypothesis ideas like comparison of score between gender, race will not affect a student's score and so on, but there is one difficulty we face to choose the suitable analysis. Most of the variables in the dataset like lunch and race are non-numeric, discontinuous and nominal variables. Those variables with no value are difficult to be used to do an accurate analysis and show relationship with the subjects' score in correlation and regression tests. To solve this problem, our group decided to combine two

related variables which are the scores of the reading and writing test, creating one new variable named language score to do a regression test with maths score.

During overcoming this problem, we have learnt a lesson that data types influence the possible analysis of data. If we want to find a relationship between two variables, either dependent or independent, numeric variables which are interval or ratio types can increase the accuracy of the result. Moreover, we have learnt on how to choose the suitable variable for each test and differentiating the differences and similarities of each test. Moreover, by doing the Data set, it has refreshed our memory about the things we learnt in Chapter 1 which are the level of measurement, continuous and discontinuous variable.

Besides that, we did face a lot of difficulties when using the RProgramming as it was our first time and with the limited time we had, it was challenging to choose the right data, method, test, obtain the correct value and finally for the plot of graphs. To overcome the challenging RProgramming, we took some time to explore and then finalize the tests that we conducted. Along the way, some changes had to be made to fit with the test and the hypothesis that we proposed. Regardless, we have successfully completed all four tests and obtained the results we expected. RProgramming has a straightforward feature to know everything within seconds, which has helped us to code our method without much hassle.

Appendix

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