

UNIVERSITI TEKNOLOGI MALAYSIA SCHOOL OF COMPUTING SESSION 2019/2020 SEMESTER 2

#### LECTURE’S NAME

DR CHAN WENG HOWE

#### PROJECT

Individual project (PROJRCT 2)

#### NAME

THANNEERMALAI A/L UDAYAPPAN

#### MATRIX NUMBER

A19EC0171

#### SECTION

02

Contents

[***Objective*** 3](#_Toc44096554)

[***Introduction*** 3](#_Toc44096555)

[***Methodology*** 3](#_Toc44096556)

[***results and discussion*** 4](#_Toc44096557)

[A. Hypothesis testing 4](#_Toc44096558)

[B. Correlation 6](#_Toc44096559)

[C. Linear Regression 9](#_Toc44096560)

[D. Chi-Square Test 13](#_Toc44096561)

[**Overall Discussion** 15](#_Toc44096562)

[E. Conclusion 16](#_Toc44096563)

[F. Acknowledgement 16](#_Toc44096564)

Student’s Performance in Exam

*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 hypothesis testing, regression model, correlation and chi-square test. Certain variables are chosen for each specific test.

# ***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 works, which is a great attribute for them to keep in mind for future class activities.

The quality of education, especially in Malaysia is becoming better and better. This implies that more effort is poured into nurturing the future generation. As a result, he test and examination being held, such as SPM, PT3, UPSR and other forms of tests are becoming progressively harder and more challenging. High school students are spending up to 2-3 hours each day in preparing for their tests or exams, while 45% of time spent studying is now allocated to test preparation. And the majority of that time is spent on tuition classes. 7 in 10 students attend tuition to cram in as much knowledge in preparation of tests.

This study is aimed to investigate the average score scored in tests among students and determine the factors that affect their marks.

# ***Methodology***

This data is retrieved from an online open source website (Kaggle) which provides and collects data. The data used in these tests are secondary data. I have decided to carry out 3 different tests on variable from the data source. From the data, I have selected 40 sample data at random and have carried Hypothesis testing, correlation test, linear regression test and chi-square test. These statistical test are done to find out the relationship between these variables. The calculations, and plots are plotted using RStudio.

# ***results and discussion***

## Hypothesis testing

Based on the previous studies conducted by Sunway on “Average Maths Scored by Students in SPM“, the survey found that the average mean scored by students are 73.66 marks out of a total 100 marks.

Hence, the null hypothesis, H0 and alternative hypothesis, H1 is:

Ho: μ = 73.66

H1: μ ≠ 73.66

Where μ is the mean of the average maths score scored by students.

A simple random sample of 40 students from secondary data source were randomly selected. They have a mean maths score of 73.6 marks. The standard deviation can be calculated by using formula:

$$s\_{x}= \sqrt{\frac{∑(x\_{i}​-xˉ)2​}{n-1}}$$

The standard deviation based on the calculation is 11.1902

A 95% level of confidence is used to test the claim of this study that the average maths score scored is not equals to 73.66 marks. The critical value of 95% level of confidence is 1.96 and -1.96. The z-value of mean can be calculated by using formula

$$z=\frac{xˉ-μ }{^{s}/\_{\sqrt{n}}} $$

And z is equal to -0.03391.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| $$xˉ$$ | μ | s | z-value | Critical value |
| 73.6 | 73.66 | 11.1902 | -0.0339 | -1.96+1.96 |

Based on the graph below, z-value=0.03391 which falls in-between our critical region as it is less than 1.96 but greater than -1.96.From the calculation and graph we fail to reject the null hypothesis, H0. This is because we have insufficient evidence to support the claim that the average maths score scored by students is not equal to 76.66 marks.

With the values from the test, I can plot a bell curve to represent my data visually (graphically) to confirm whether my test statistic fall in the critical region or not.

Graph 1

## Correlation



Graph 2

In this correlation test, I am analysing the relationship (means of strength of association) between reading scores and writing scores obtained by a random sample size of 40 students.

The coefficient correlation, r is calculated using SPSS to show the relationship between these two variables. Since the data being tested is in ratio form, hence the coefficient correlation is calculated using the Pearson’s technique and can be referred as bivariate correlation.

The independent variable(X) is the reading score and the dependant variable(Y) is the writing score. These two variables are used to calculate the coefficient correlation using the formula below.



The correlation coefficient (r) is 0.9339426, which shows a strong positive relationship between these two variables. As an inference from the test result, as the writing score increases the reading score also increases. Based on the scattered plot below, it shows a positive slope.

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

Significance Test for Correlation

Then, these two variables are used to test whether there is any evidence of linear relationship between them at 0.05 level of significance. H0 is assumed that the reading score has no correlation

Ho: ρ = 0; p = population correlation coefficient

H1: ρ ≠ 0

In this test, the test statistics,$t= \frac{r\sqrt{n-2}}{\sqrt{1-r^{2}}}$ where sample size n = 40 and r equals to 0.9339426 as calculated above. Using the formula above, value of t is equivalent to 16.108, with degree of freedom of 38. The critical value is found out to be -2.024 and 2.024.





Graph 3

T0=16.108

From the graph above, it can be concluded that the test statistic fall in the critical region .Since the test statistic, t (16.108) is greater than critical value (2.024). Hence we will reject the null hypothesis at α = 0.05. There is sufficient evidence to claim that there is a linear relationship between the reading score and writing score at a 95% confidence level. We also have sufficient proof to conclude that the population correlation coefficient, P is not equal to zero,

## C:\Users\Mala\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Rplot06.png Linear Regression

Graph 4

Linear regression is a linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables).Linear regression is one of the most commonly used predictive modelling techniques. The aim of linear regression is to find the relationship for a continuous response variable Y as a function of one X variable. So that this regression model to predict the Y when only the X is known.

This mathematical equation can be generalized as follows:

Y = β0 + β1X + ε

Y is dependent variable,

β0 is the population intercept y,

β1 is the population slope coefficient

X is independent variable,

ε is random error component.

This regression show the result of the relationship between reading and writing scores (marks).

Dependent variable = Reading score

Independent variable = Writing score

From the RStudio calculation above, the estimate of regression intercept is 5.5995(b0) and the estimated regression slope is 0.9037.

The estimated regression model is then calculated using

 ŷ = 5.5995 + 0.9037x

Moreover, the coefficient of determination, R2 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 (R2) is then calculated using RStudio in which we obtain, R2 = 0.8722 which show that only 87.22 % of the variation in reading score is explained by the writing score.

The next step is to calculate the test statistic:

The following formula is used:



The b1=0.9037, is from the previous calculation. Sb1 is the estimator of the standard error of the slope.

The following formula is used:



Se is calculated using the following formula:



SSE is calculated using the following formula:



Se=4.149 and the value of Sb1=0.0561

The test statistic is than calculated, T=16.1087

+T (38, 0.025) = 2.0244

- T (38, 0.025) = -2.0244

df= 38

As observed from the calculations above, the test statistic is greater than the critical region as T=16.1087 > T (38, 0.025) = 2.0244. This means we reject the null hypothesis. That means that there are sufficient evidence to prove that there is a linear relationship between reading and writing score at a 95% confidence level. Hence the writing score can be predict using the reading score and vice versa.

## Chi-Square Test

A Chi-Square test was performed to test how likely it is that the observed distribution is due to change. The test also measures weather the observed distribution of data fits with the expected distribution. I have used the variables of gender and test preparedness to test whether these variables are independent to one another at a significance level of 0.05.The observed values are then compared to the corresponding expected values. The null hypothesis, Ho and alternative hypothesis, H1 is as follows:

Ho: There is no relationship between gender and test preparedness and both variables are independent to one another.

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

The following code was used to obtain the test statistics for Chi-Square test.







The observed values and expected values are as shown as above. The test statistic, chi-square value is then calculated using RStudio which equals to **0.074098**. The critical value can then be found based on the Chi-Square Distribution table, with degrees of freedom equals to 1 and significance level of 0.05. The p- value is found to be approximately 0.7855.

From the calculation, we can see that the test statistic does not fall in the critical region.

X2=0.7855 which is less that X2 (1, 0.05) = 3.841, thus we fail to reject the null hypothesis. I can conclude that at a 95% confidence interval there is sufficient evidence to prove the relationship between gender and test preparedness.

# **Overall Discussion**

Base on the statistical tests I have conducted, I can now interpret the result as whole and be able to discuss the relationship between variables. Firstly, it is stated that the average maths score, scored by students in a maths test is 73.66. That means an average students test score maths exam is 73.66 and a student will be roughly be able to estimate his/her test preparedness level. Next, there is a confirmed relationship between the reading score and writing scores. This is supported due to the outcome of the correlation test carried out. The reading score represents a student’s fluency in speaking a certain language. And the writing score represents how good a students is able to express his/her idea on a piece of paper. This result is necessary to the students because it is found out that both these variables are strongly related. So when one score increases the other is also most likely to increase and vice versa.

Students can use this knowledge to prepare for upcoming tests and take advantage of this. Instead of preparing for two separate tests, they would benefit more focusing on 1 test instead of both. This will help the students gain better marks in both reading and writing scores in their tests. By carrying out linear regression analysis, we now have a certain answer that there is a relationship between reading scores and writing scores. From this test, we are also able to predict the score that the students is most likely to obtain based on one of the test scores know. This prediction is likely to be accurate due to the strong positive relationship of these variables.

Lastly, by conducting the chi-square test of independence, it gives us a decisive factor contributing to this case study. Gender plays an important role when it comes to test preparedness. Different genders have a different take on preparing for a test. One gender is more likely to prepare for a test when compared to another. This is reflected in their test scores obtained from 3 different tests.

## Conclusion

 Based on the hypothesis, we fail to reject the null hypothesis. There is insufficient evidence to support the claim that the average math’s score scored by students is not equal to 76.66 marks.

 Next, from the analysis, 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.

 The estimated regression model is then produced in which we obtain ŷ = 5.5995 + 0.9037x, and this regression model is helpful in predicting the average reading score based on average writing score.

 The conclusion which can be drawn from the chi square test is since the p-value is > 0.05, therefore fail reject the null hypothesis. As a conclusion, there is evidence of a relationship between gender and test preparedness.

* 1. Acknowledgement

 I would like to extend my sincerest apperception to our bellowed lecturer, Dr Chan Weng Howe for guiding me throughout the entirety of this project

G. Reference

1. Correlation Test between Two Variables in R. (n.d.). Retrieved June 25, 2020, from

2. Richie Cotton Richie (1961, January 01). Linear regression in R (normal and logarithmic data). Retrieved June 25, 2020, from https://stackoverflow.com/questions/6277561/linear-regression-in-r-normal-and-logarithmic-data

3. Chi-squared Test of Independence. (n.d.). Retrieved June 25, 2020, from http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

4. Robk@statmethods.net, R. (n.d.). Probability Plots. Retrieved June 25, 2020, from https://www.statmethods.net/advgraphs/probability.html