UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 - SECTION 04
PROBABILITY AND STATISTICAL DATA ANALYSIS

PROJECT 2:
A STUDY TO UNDERSTAND THE INFLUENCE OF EXTERNAL FACTORS TO THE STUDENTS' PERFORMANCE IN EXAMS

VIDEO LINK: https://youtu.be/b1UmXPV8cyA

LECTURER'S NAME: DR. ARYATI BINTI BAKRI

N0.	**NAME**	**MATRIC NUMBER**
1.	MUHAMMAD NUR SOLIHIN BIN MALIK RADZUAN	A21EC0089
2.	SARAH FARHANA BINTI SALLEH	A21EC0226
3.	NORAIN BINTI MOHD SULAIMAN	A21EC0106
4.	MOHAMAD SYAFIQ FIRDAUS BIN ABDUL AZIZ	A21EC0055

TABLE OF CONTENTS

INTRODUCTION

The dataset was already prepared by our lecturer who is Dr. Aryati Bakri. This dataset presented all the variables that explained the student performance based on 3 scores which are writing score, mathematics score, and reading score. This data set consists of the marks secured by the students in various subjects. Each variable that is included in the dataset is meaningful and suitable for us to do an analysis. We chose to do a few tests on the variables that are represented in the dataset. Firstly, we make the one-sample hypothesis testing to test the average score of 3 scores for both genders which are men and women. This hypothesis testing is to know whether it is true or not the average score for 3 tests that have been done by the students above 65 and 50 respectively. To test that claim, we do the one-sample hypothesis testing on a few variables which are gender and the average score of 3 tests. Secondly, we make a correlation analysis to investigate the relationship between the Mathematics score and Writing score. This analysis is done in order to know whether there is a correlation between both variables. Thirdly, we make a regression analysis to investigate the relationship between reading scores and writing scores. This analysis was done in order to know whether the writing score is affected by the reading score if the reading score changes. In order to know that relationship, regression analysis is the best technique because we can test between the independent variable and the dependent variable. Lastly, there is a chi-square test of independence to determine if there is a significant relationship between gender and the test preparation course. It can be said that we want to test whether gender plays an important role or not for the students to complete their test preparation course. Hence, the chi-square test of independence is a suitable method to know that kind of claim. In conclusion, all of the tests that have been conducted by each of the group members on each variable that we gain from the dataset that has been provided.

Test 1: One sample test to test the average score for 3 tests (male)

This one sample testing is to test whether the statement is true that the average score for 3 tests of male students is above 65. Assume the confidence level to be 95% and the significance level, α = 0.05. Let the population average score for 3 tests of male students be μ.

H0: μ = 65
H1: μ > 65

Calculations

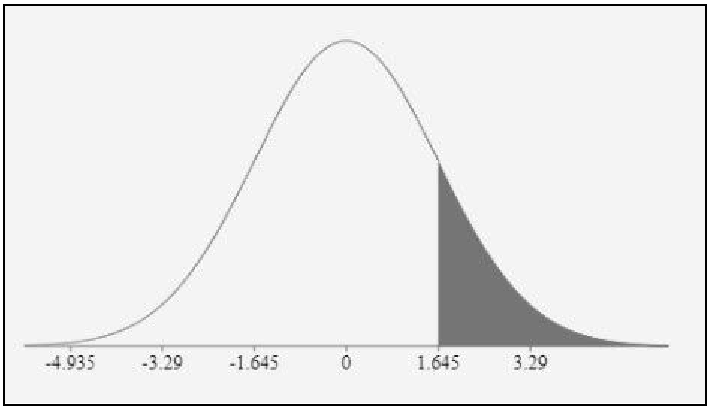
Based on the analysis of the R programming, all the values have been summarized into the table.

α	0.05
Sample size , n	482
Sample mean , x̅	65.837
Sample Standard deviation, σ	13.699

Test Statistic

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

Test statistic , z	1.342
Critical value , c.v = Z0.05	1.645



The shaded region is the rejection region.

Decision

Since the test statistic value = 1.342 is smaller than the critical value = 1.645, which falls outside the rejection region, therefore we fail to reject the null hypothesis.

Conclusion

There is insufficient evidence to support the statement that the average score for 3 tests of male students is above 69. In other words, the average score of males is 65.

Test 2 : One sample test to test the average score for 3 tests (female)

This one sample testing is to test whether the statement is true that the average score for 3 tests of female students is above 50. Assume the confidence level to be 95% and the significance level, α = 0.05. Let the population average score for 3 tests of male students be μ.

H0: μ = 50
H1: μ > 50

Calculations

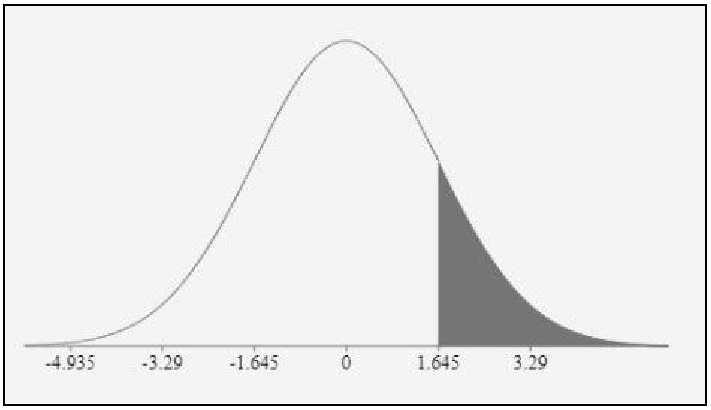
Based on the analysis of the R programming, all the values have been summarized in the table.

α	0.05
Sample size , n	518
Sample mean , $\bar{x}$	69.569
Sample Standard deviation, σ	14.542

Test Statistic

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

Test statistic , z	30.629
Critical value , c.v = Z0.05	1.645



The shaded region is the rejection region.

Decision

Since the test statistic value = 30.629 is greater than the critical value = 1.645, which falls inside the rejection region, therefore we reject the null hypothesis.

Conclusion

There is sufficient evidence to support the statement that the average score for 3 tests of female students is above 50.

Test 3: Correlation Analysis to investigate the relationship between the Mathematics score and Writing score.

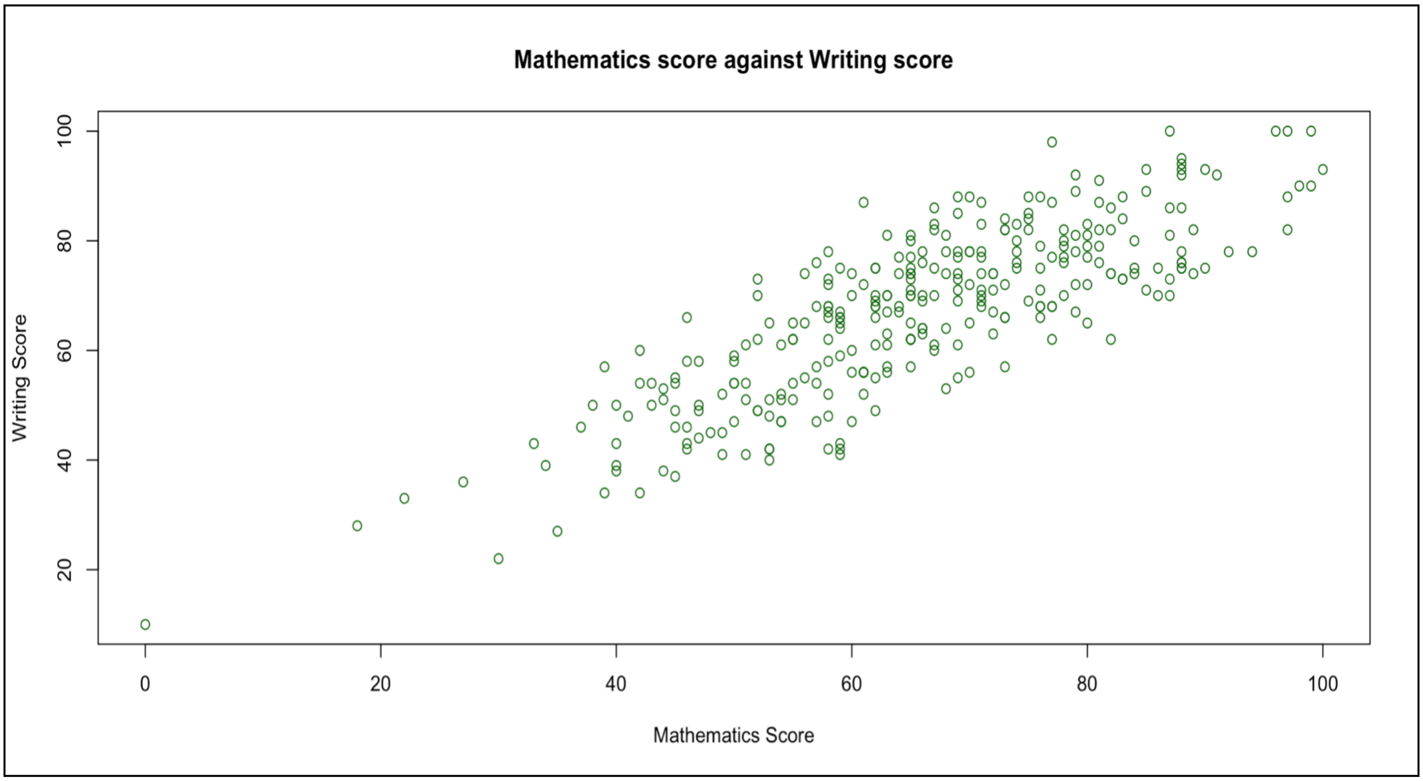


Figure 3: Writing Score and Mathematics Score

This test is to measure the strength of the relationship between Mathematics score and Writing score.
Assume the confidence level to be 95%, significant level, α = 0.05.

H0: ρ = 0 (no linear correlation between Mathematics score and Writing score)

H1: ρ ≠ 0 (linear correlation exists between Mathematics score and Writing score)

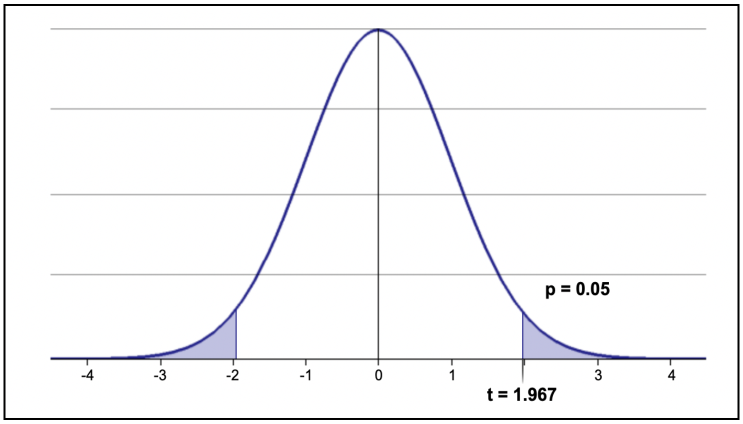
Calculations

Based on the analysis of the R programming, all the values have been summarized into the table.

α	0.05
Correlation coefficient,r	0.8235336
Sample size , n	300
Degree of freedom , df	298

Test Statistic

Degree of freedom, df = 298
Critical Value,



The shaded region is the rejection region.

Critical value , t (0.025, 298)	1.967
Critical value , -t (0.025, 298)	- 1.967

The critical points shown in the t distribution are 1.967 and -1.967

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

Test statistic, t = 25.061

Decision

Since the test statistic, t = 25.061 is greater than t (0.025, 298) = 1.967 and -t (0.025, 298) = -1.967. It falls within the rejection area. Hence, we reject the null hypothesis that says there is no linear correlation between Mathematics score and Writing score.

Conclusion

The correlation coefficient, r = 0.8235336 is a positive value and falls within 0.8 and 1. Hence, it has a strong linear relationship between Mathematics scores and Writing scores. There is sufficient evidence to prove that a linear correlation does exist between Mathematics score and Writing score

Test 4: Regression Analysis to investigate the relationship between Reading score and Writing score

Assume that the confidence level is 95%, significant level, α = 0.05

H0: β1 = 0 (no linear regression between Writing score and Reading score)

H1: β1 ≠ 0 (linear regression exists between Writing score with Reading score)

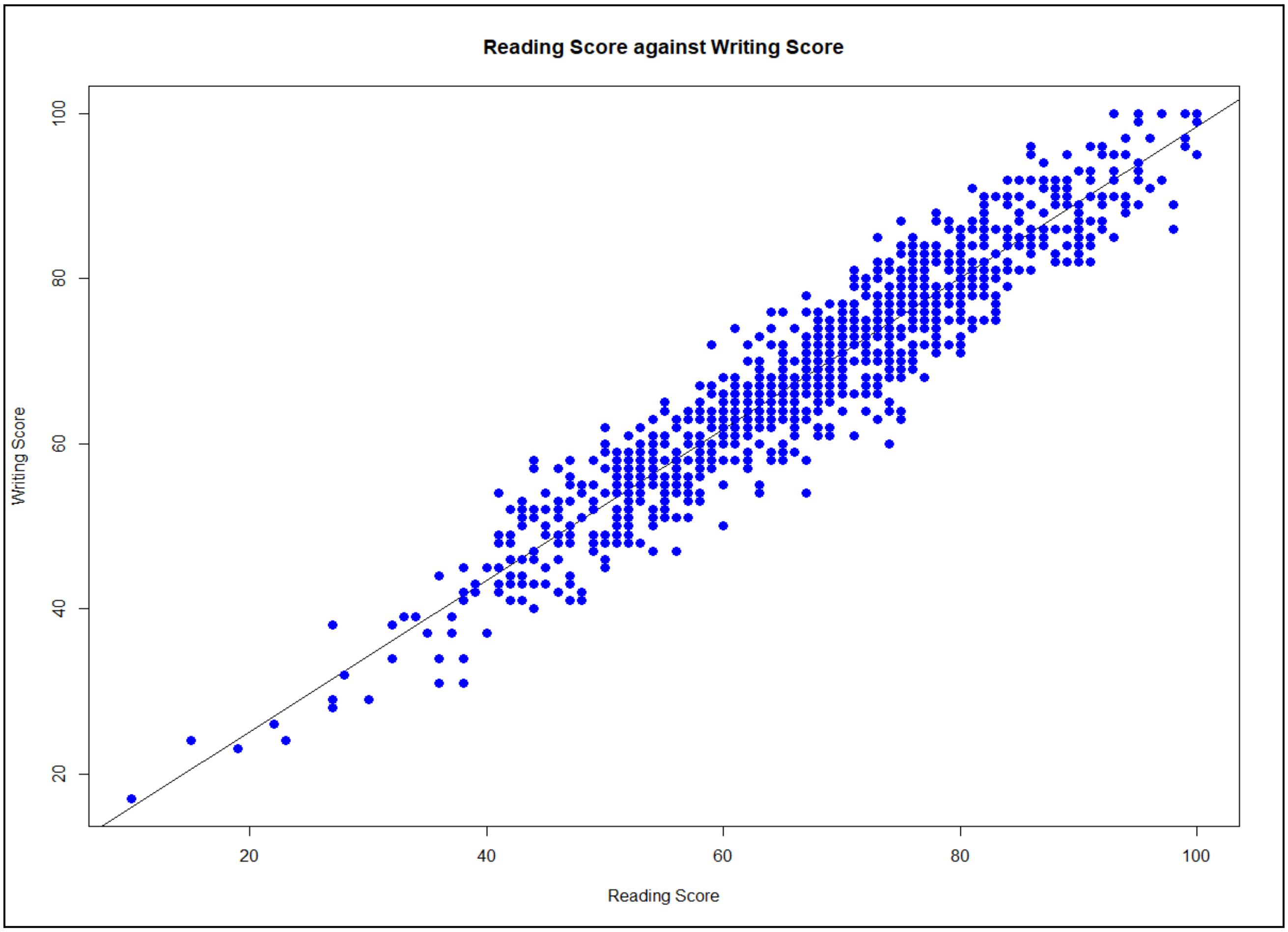


Figure 3: Writing Score and Reading Score

The independent variable (a variable that is used to explain the dependent variable) is Reading score while the dependent variable (a variable that we wish to explain) is Writing score.

The r-squared value of regression R^2 is 0.9113.

This shows that there is 91.13% of the variation in writing score is explained by the reading score.

Calculations

Based on the analysis of the R programming, all the values have been summarized in the table.

α	0.05
Sample size , n	1000
Degree of freedom , df	998
Regression Line	ŷ = -0.6675 + 0.9935x

Test Statistic

Degree of freedom, df = 998

Critical Value,

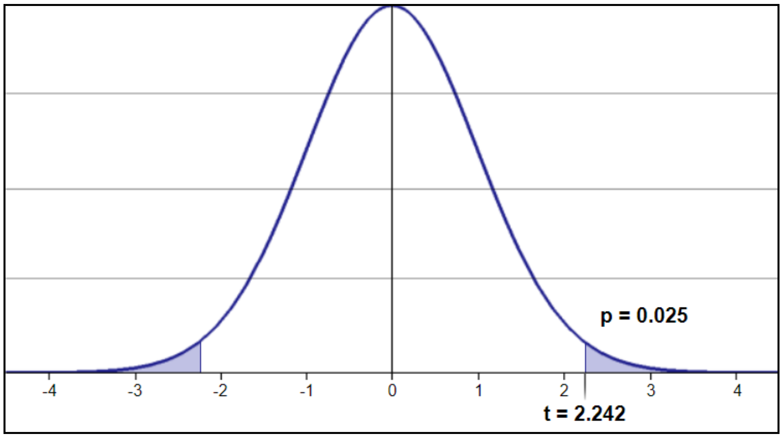


Figure 4 : Normal distribution of t value

Range	T- value
-t(0.025,998)	-2.242
t(0.025,998)	2.242

The shaded region is the rejection area

The critical points shown in the normal distribution are -2.242 and 2.242.

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

Test Statistic,
t = 101.233

Decision

Since the test statistic, t = 101.233 is greater than t(0.025,998) = 2.242 and -t(0.025,998) = -2.242. It falls within the rejection area. Hence, we reject the null hypothesis saying that there is no linear regression between writing score and reading score.

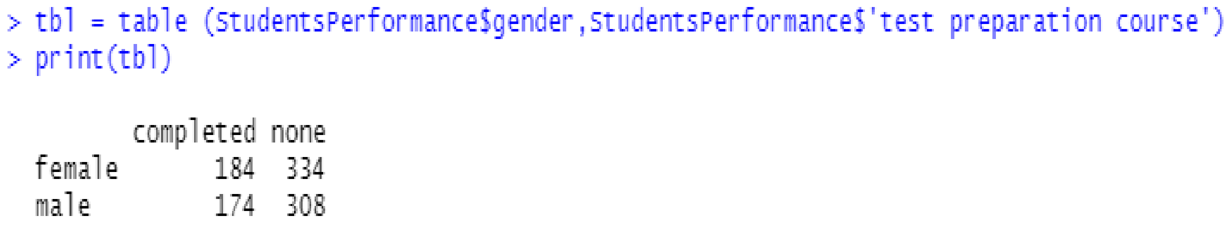
Conclusion

There is sufficient evidence that linear regression exists between the writing score and the reading score. Hence, there is enough evidence that reading scores affect writing scores.

Test 5: Chi-Square Test of independence to determine whether there is a significant relationship between gender and test preparation course

Assume the confidence level to be 95%, significant level, α = 0.05.

In this analysis, we are using the variable gender of the students and the test preparation course, where we will test whether the number of doors and car aspiration are related using the Two Way Contingency Table, at a 95% confidence level. Hence, we use the Chi-Square Test of Independence, with a two-way contingency table.



```
> tbl = table (StudentsPerformance$gender,StudentsPerformance$'test preparation course')
> print(tbl)

        completed none
 female       184  334
 male         174  308
```

Observed frequencies for the gender and the test preparation course.

Contingency table obtained in Rstudio

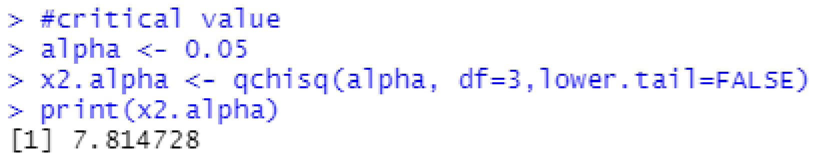
Gender	Test Preparation Course	
	completed	none
Male	184	334
Female	174	308

Test Hypothesis

H0: The gender of the students and the test preparation course are independent.

H1: The gender of the students and the test preparation course are dependent.

Critical value



```
> #critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=3,lower.tail=FALSE)
> print(x2.alpha)
[1] 7.814728
```

Finding critical value x^2 using RStudio

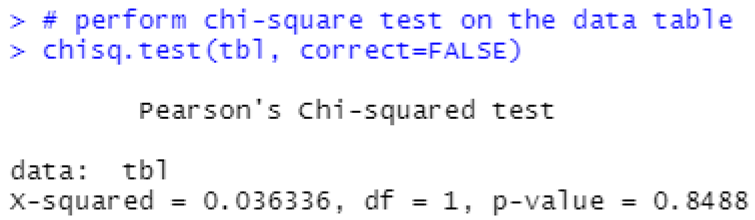
Critical value x^2 = 7.815 (with df=(2-1)(2-1)=1, α = 0.05)

Calculate expected value

Gender	Test Preparation Course				Total
	completed		none		
	Observed	Expected	Observed	Expected	
Male	184	$\frac{358\ X\ 518}{1000}$ =185.44	334	$\frac{642\ X\ 518}{1000}$ =332.56	518
Female	174	$\frac{358\ X\ 482}{1000}$ =172.56	308	$\frac{642\ X\ 482}{1000}$ =309.44	482
Total	358	358	642	642	1000

*Remarks: $e_{ij} \geq$ in all cells

Test statistic using RStudio



```
> # perform chi-square test on the data table
> chisq.test(tbl, correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 0.036336, df = 1, p-value = 0.8488
```

When we calculate test statistics using RStudio, we get test statistic x^2 = 0.036336, with *p-value* = 0.8488.

Decision

Since the test statistic value (x^2 = 0.036336) < critical value ($x^2_{k=1,\ \alpha\ =\ 0.05}$ = 7.815), it does not fall within the critical region. Thus, we fail to reject H0. There is sufficient evidence to conclude that there is an independent relationship between the variables gender of the students and the test preparation course, at α = 0.05.

CONCLUSION

In conclusion, the statement that claims the average score for 3 tests of male students is above 65 is false since the result of the hypothesis test, we fail to reject the null hypothesis, hence we can say that the average score for 3 tests of male students is equal to 65 as there is no enough evidence to prove otherwise. Next, the statement that states the average score for 3 tests of female students is above 50 is true. The result of the test came out that we can reject the null hypothesis, which is that the average score for 3 tests is equal to 50. There is enough evidence to prove that the average score for 3 tests for female students is above 50. Next, there is a strong linear relationship between Mathematics scores and writing scores. Other than that, there is a linear regression between writing score and reading scores. Lastly, it has been proved that the gender of the students and the test preparation course are dependent.

REFERENCES

- Example, O. and SPSS, H., 2022. *One-Sample T-Test: SPSS, By Hand, Step by Step*. [online] Statistics How To. Available at: <https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-sample-t-test/> [Accessed 27 June 2022].

- *Regression Analysis*. Corporate Finance Institute. (2022). Retrieved 28 June 2022, from https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/.

- Toolpak, H. (2022). *Correlation in Statistics: Correlation Analysis Explained*. Statistics How To. Retrieved 28 June 2022, from https://www.statisticshowto.com/probability-and-statistics/correlation-analysis/.