



**SECI2143**





**PROJECT 2**

**DATA ANALYSIS (INFERENTIAL STATISTICS)**

**Student's Enrollment To College**

**SECTION 6 - GROUP 10**

**Lecturer : Dr. Aryati Binti Bakri**

				
NAME	MUHAMMAD NAJWAN HAZIM BIN KHAIRI	MUHAMMAD TAUFIQ BIN JURIMI	AFIQ FAHMI BIN ROSLAN	AZHAN HANIFF BIN AZNI
STUDENT ID	A21EC0087	A21EC0095	A21EC0153	A21EC0017

<b>CONTENT</b>	<b>PAGE</b>
INTRODUCTION	1
DATASET	2
DATA ANALYSIS	3
One Sample Hypothesis Testing	3
Regression test	6
Chi Square	8
CONCLUSION	9
APPENDIX	10

## INTRODUCTION

Nowadays, knowledge is very important for students as it sharpens skills like reasoning and problem-solving. Student interest refers to the inclination of the student towards a particular subject in which he or she is easily able to connect without any hassle or hurdle. However, student interest towards studies declines day by day. School counselors are responsible for helping students by finding the factors and helping them. These factors can be observed in many ways. For example, these students can be judged in terms of its specifications such as type school, school accreditation, gender, interest, residence, parent age and many more. To accomplish this goal, a few variables have been chosen and some test studies are conducted.

The objective of this project is to gain knowledge on how to use RStudio and R programming language for statistical purposes. We can use this dataset to discover and describe the pattern or relationship between different variables. With the help of a sample size of 1000 respondents, we can use the dataset to draw inferences and identify some characteristic or overarching pattern regarding a population.

## DATASET

The data in this study is obtained through secondary data collection, which is entitled “Go To College Dataset”. This dataset aims to predict whether different factors affect students who will continue to go to college or not. By identifying the cause and offering assistance, school counselors can assist students who won't attend college.

The data which are respective to the 1000 respondents are selected from the database and used for hypothesis testing to determine whether there is enough evidence to support the null hypothesis for hypothesis testing, correlation, regression and chi square. The sample is normally distributed and plotted with RStudio.

Variables	Type of Variable	Level Of Measurement
type_school	Qualitative	Nominal
school_accreditation	Qualitative	Nominal
gender	Qualitative	Nominal
interest	Qualitative	Ordinal
residence	Qualitative	Nominal
parent_age	Quantitative	Ratio
parent_salary	Quantitative	Ratio
house_area	Quantitative	Ratio
average_grades	Quantitative	Interval
parent_was_in_college	Qualitative	Nominal
in_college	Qualitative	Nominal

## DATA ANALYSIS

### One Sample Hypothesis Testing

School counselors conducted a survey to help students that will not go to college by finding the factor. A random sample of 1000 respondents is obtained, and each person's average grade is measured. A 0.05 significance level is used to test the claim that the mean sample is greater than population mean which is 84, which is a value often used for the upper limit of range of normal values.

#### 1. Hypothesis Statement :

- $H_0 = \mu = 84$
- $H_1 = \mu > 84$

#### 2. Given 95% confidence level, $\alpha = 0.05$ . The z test statistics, $Z_0$ can be calculated by :

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- $\bar{X} = 86.0972$
- $\mu = 84$
- $S = 3.378738$
- $n = 1000$
- By using RStudio, test statistics,  $Z_0 = 19.6284195723525$

#### 3. Calculate the critical value :

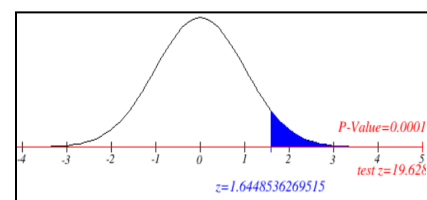
- By using RStudio, critical values,  $Z_\alpha = 1.64485362695147$

#### 4. Decision :

- $H_0$  is rejected

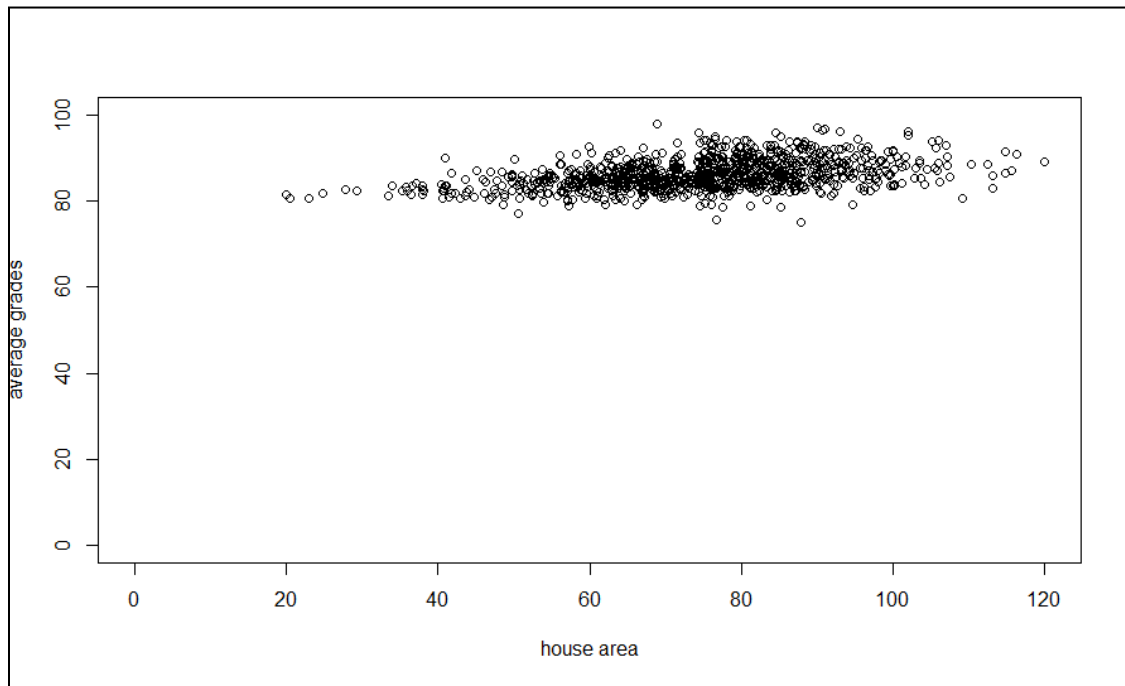
#### 5. Conclusion :

- Since  $Z_0 = 19.628 > Z_\alpha = 1.6448$ , we reject  $H_0$ , there is sufficient evidence that the mean of grade exceeds 84.



## Correlation test

For correlation, we want to observe if there is a linear correlation between **house area**(house\_area) and **average grades**(average\_grades). To test this hypothesis, we used Pearson's Product-Moment Correlation Coefficient, at 95% level of confidence.



Based on the scatter plot above, it is observed that there is a slight positive correlation relationship between **house area** and **average grades**. We can make a conclusion that when house area increases the average grades also increases.

### 1. Calculate the Sample Correlation Coefficient (r)

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

[Pearson's method formula to find the sample correlation coefficient (r)]

**house\_area = x, average\_grades = y**

With the help of RStudio, we obtained a sample correlation coefficient,  $r = 0.4095654$ , which shows that x and y have a reasonably weak positive linear correlation.

## 2. Significance Test for Correlation

Hypothesis Statement :

- $H_0 = \rho = 0$  (no linear correlation)
- $H_1 = \rho \neq 0$  (linear correlation exists)

3. Given 95% confidence level,  $\alpha = 0.05$ . The test statistics, t can be calculated by :

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

With the help of RStudio, we obtained the test statistic,  $t = 14.1827$

## 4. Calculate the Critical Value

$\alpha = 0.05$ ,  $df = n-2 = 998$

There are two critical values since this is a two-tailed test, by referring to the t-table we acquired the values:

Lower tail critical value  $-t_{(\alpha/2 = 0.025, df = 998)} = -1.9623$

Upper tail critical value  $t_{(\alpha/2 = 0.025, df = 998)} = 1.9623$

## 5. Decision

- Since test statistics  $t = 14.1827 > \text{Upper Tail Critical Value } t_{(\alpha/2 = 0.025, df = 998)} = 1.9623$
- $H_0$  is rejected

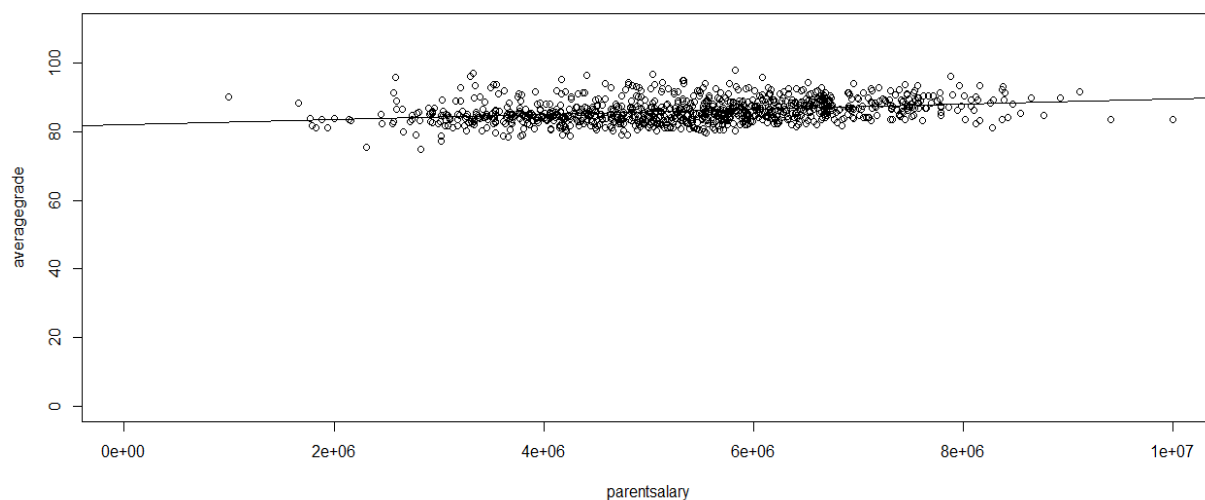
## 6. Conclusion

- There is sufficient evidence that supports  $H_1$  that claims that there exists a linear correlation between **house area(house\_area)** and **average grades(average\_grades)**.

## Regression test

For regression, we want to see whether there is a relationship between parent salary and average grade. The independent variable is parent salary while the dependent variable that we want to explain is average grade.

### PARENT SALARY AGAINST AVERAGE GRADE



#### 1. Regression test explanation

$$\hat{y} = 82.11 + 0.0000007415x$$

y-Intercept = 82.11

Slope = 0.0000007415

Intersection Coefficient,  $b_0 = 82.11$

Interpretation of the Slope Coefficient,  $b_1 = 0.0000007415$

$b_0$  is the estimated average value of average grade when the parent's salary is zero. But here, there is no parent with 0 salary so  $b_0$  just indicates that for average grade within the range of parent salary, 82.11 is the portion of the grade not explained by parent salary.

$b_1$  tells us that the average value of parent salary increases by 0.0000007415 on average, for each additional amount of salary.



$$R = 0.3067119$$

It can be seen that average grade increases as the parent's salary increases, indicating that there is a positive relationship between parent salary and average grade.

$$R^2 = 0.09407$$

This shows that there is a weak linear relationship between average grade and parent salary. Only 9.407% of the average grade achieved is explained by the parent's salary.

## 2. Inference about the Slope: t Test

Muhammad Taufiq wants to test the claim that there is no regression relationship between parent salary and average grade, at 95% of confidence level.

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

$$\alpha = 0.05, n = 1000, \hat{y} = 82.11 + 0.0000007415x, df = 998$$

Critical value,

$$-t_{0.025, 998} = -1.962$$

$$t_{0.025, 998} = 1.962$$

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

By using RStudio,

$$s_{b_1} = 0.00000007284$$

$$t = 10.18$$

## 3. Decision:

Since  $t_0 = 10.18 > t_{\alpha} = 1.962$ , we reject  $H_0$

## 4. Conclusion:

There is enough evidence that a linear regression exists between parent salary and average grade.

There is sufficient evidence that parent salary affects students' average grade.

## Chi Square

For chi square, We're looking to see if there's a connection between two qualitative variables which is between type of school (type\_school) and residence. 1000 data sample is used and the type of school is divided into two groups which are academic and vocational while residence is divided into two groups which are rural and urban. We claim that, at 95% of significance level, the type of school is independent of the residence.

$H_0$  : the type of school is independent to the residence.

$H_1$  : the type of school is dependent to the residence.

We are using this formula to calculate expected count  $e_{ij}$  :

$$e_{ij} = \frac{(i^{th} \text{ Row total}) (j^{th} \text{ Column total})}{\text{Total sample size}}$$

And for test statistic,  $\chi^2$  we are using this formula and apply it into Rstudio :

$$\chi^2 = \frac{\sum (O_{ij} - E_{ij})^2}{E_{ij}}$$

	Residence		
Type of school	Rural	Urban	Total
Academic	221	388	609
Vocational	240	151	391
Total	461	539	1000

From the Rstudio, we obtain the  $\chi^2 = 60.336$  and the degree of freedom,  $df = 1$  . The critical region can be found in the Chi-Square Distribution table with a significance level of 0.05 and a degree of freedom of 1. We found that the critical value is approximately 3.84145882069413 and the p-value is approximately 0.000000000000007997.

In conclusion, we observe that test statistic,  $\chi^2 >$  critical value which is  $60.336 > 3.84145882069413$ . Test statistics fall within the critical value, thus we reject the null hypothesis. There is enough evidence that the type of school is dependent to the residence.

## CONCLUSION

One sample hypothesis testing has been conducted, we test on the mean and the variances are unequal. The testing has shown that the mean of student average grade which is 86.09 is greater than the population mean of student average grade which is 84.0 , hence we reject the null hypothesis.

Furthermore, when conducting the correlation analysis, we tested if there is a linear correlation between house area and average grades, we can conclude that there is a linear correlation between those two data, hence we reject our null hypothesis. The correlation indicates a reasonably weak linear correlation, because our sample correlation coefficient  $r = 0.4095654$ .

Moreover, from the regression analysis, we are 95% confident that parent salary positively affects students' average grade. The relationship between parent salary and student's grade is a weak positive linear relationship, and only 9.407% of the average grade achieved is explained by the parent's salary.

Finally, from the chi-square calculation and analysis, We are 95% certain that the type of school is related to the residence. We reject the null hypothesis because the test statistic is within the critical value or we can say that the p-value is lower than significance level.

In conclusion, as we can see we can perform many test statistics tests such as one sample test hypothesis testing, correlation analysis, regression analysis and chi square test by using R studio. We would like to thank Dr Aryati for all the knowledge given to us. This project has improved our skill in R programming which is important for our use in the future.

PRESENTATION VIDEO:

<https://drive.google.com/drive/folders/1qMWAKDcJayBaN9Z9p6vvJeRGjQtgAM-Q?usp=sharing>

DATASET:

<https://docs.google.com/spreadsheets/d/1r3JMWlcant9epiSd1onQtf-F7KuvkieXhW0JULCXEo/edit?usp=sharing>

EPORTFOLIO REFLECTION:

NAME	STUDENT ID	EPORTFOLIO LINK
MUHAMMAD NAJWAN HAZIM BIN KHAIRI	A21EC0087	<a href="https://eportfolio.utm.my/view/view.php?t=8aBIFv4ANqWzkjE5lYd9">https://eportfolio.utm.my/view/view.php?t=8aBIFv4ANqWzkjE5lYd9</a>
AFIQ FAHMI BIN ROSLAN	A21EC0153	<a href="https://eportfolio.utm.my/view/view.php?t=MoUt0dDz1L7q9pVxP3Th">https://eportfolio.utm.my/view/view.php?t=MoUt0dDz1L7q9pVxP3Th</a>
AZHAN HANIFF BIN AZNI	A21EC0017	<a href="https://eportfolio.utm.my/view/view.php?t=xEIU09pFO1g8hBfciaYI">https://eportfolio.utm.my/view/view.php?t=xEIU09pFO1g8hBfciaYI</a>
MUHAMMAD TAUFIQ BIN JURIMI	A21EC0095	<a href="https://eportfolio.utm.my/user/muhammad-taufiq-bin-jurimi/seci2143-06-kebarangkalian-statistik- analisis-data-probability-statistical-data-analysis">https://eportfolio.utm.my/user/muhammad-taufiq-bin-jurimi/seci2143-06-kebarangkalian-statistik- analisis-data-probability-statistical-data-analysis</a>

## **APPENDIX**

P. Roxy, O. Chris, D. Jay. (2012). Introduction to Statistics & Data Analysis Fourth Edition.

Richard Stratton

Go To College Dataset. (n.d.). Retrieved from

<https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset>