UNIVERSITI TEKNOLOGI MALAYSIA

SESSION 2021/2022, SEMESTER 2

SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

PREPARED FOR:

DR. ARYATI BINTI BAKRI

SUBMISSION DATE:

3 JULY 2022

Prepared By:

GROUP 5			
NO.	**NAME**	**MATRIC NO**	**SECTION**
1.	MUHAMAD FAIZ BIN ABDUL MUTALIB	A21EC0059	04
2.	MUHAMMAD QAYYIM BIN KHAMARUDIN	A21EC0090	
3.	PUAH JUN HONG	A21EC0221	
4.	YAW CHOON HONG	A21EC0240	

Presentation Video Link: https://drive.google.com/file/d/1XjGOP-LB5txkKcd5sz3a-rW-9Q_iRFDU/view?usp=sharing

Table of Contents

1. Introduction

The dataset that involved in this project is about the decision of the Indonesian high school students to go to the college in Indonesia and it was originally created by Saddam Sinatrya Jalu Mukti (Kaggle, 2022). Our team has obtained the dataset from Kaggle.

In this project, our team would like to use the obtained dataset to investigate and understand the factors that will be affecting the Indonesian high school students to decide their decision to go to the college. Since the dataset consists huge number of records of the students' responses, hence our team would like to utilize the RStudio Statistical Tool, so that the inferential statistics process will be conducted more efficient and easier.

To involve the inferential statistics into this project, our team has implemented 4 types of tests which are the one sample hypothesis testing, Chi-Square test, correlation and regression. To explain about the testing that was implemented, one sample hypothesis testing will be used to observe whether the parent's salary of the students is above the average monthly salary of the ordinary workers in Indonesia. Next, Chi-Square test will be used to determine whether the school accreditation and type of school are having independence relationship with each other. Furthermore, correlation testing will be used to observe the existence of linear correlation between the parent salary and the students' average grades. Last but not least, the regression testing will be used to determine the existence of linear regression between the students' house area and the students' average grades.

2. One Sample Hypothesis Test

According to Statistics Indonesia (June, 2021), it stated that the average monthly income of the worker in Indonesia is IDR 2,756,350. Conduct a hypothesis test on the salary of the Indonesian parents to determine whether the Indonesia parents has a higher average monthly income than the ordinary Indonesian worker or not. At significance level = 0.05, is there are enough evidence to support the statement above?

H0: μ = IDR 2756350

H1: μ > IDR 2756350

Alpha = 0.05

Critical Value = 1.6448536

Sample Size: 1000

Sample Mean: 5381570

Sample Standard Deviation: 1397545.90968228

Test Statistics: $\boldsymbol{Z} = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ = 59.401802126948

Decision:

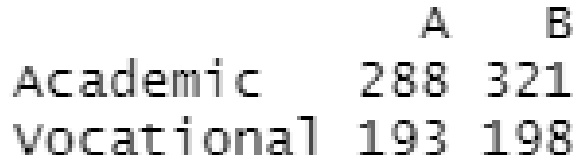
Since Z-test statistics > Z-critical value, 59.401802126948 > 1.64485362695147. Hence, reject Ho.

Conclusion:

There is sufficient evidence show that the Indonesia parents from the sample have a higher average monthly income than the ordinary Indonesian worker.

3. Chi-Square Test of Independence

Chi Square test of independence to determine if there is a significant relationship between the school accreditation and the type of school.



	A	B
Academic	288	321
Vocational	193	198

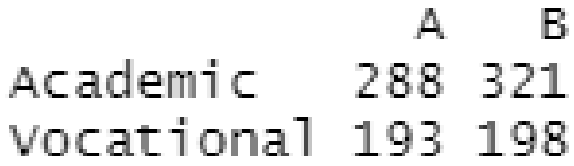
Figure 3.1: Contingency table of the data after inputting in RStudio

Assume the confidence level to be at 95%, which the significant level, $\alpha = 0.05$.

Hypothesis stating:

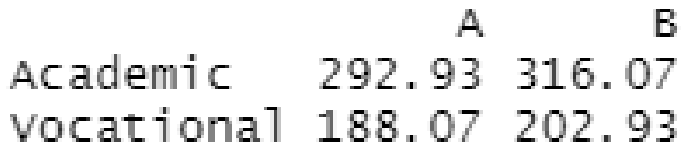
H_0: The school accreditation and the type of school are independent.

H_1: The school accreditation and the type of school are related and is dependent.



	A	B
Academic	288	321
Vocational	193	198

Figure 3.2: Observed count contingency table



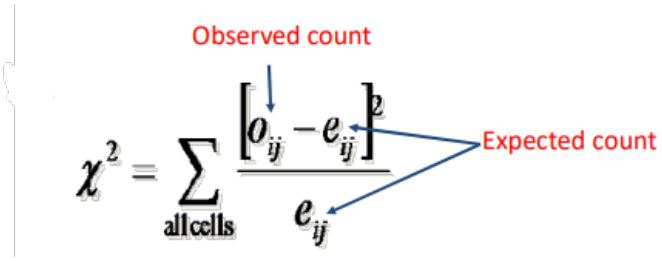
	A	B
Academic	292.93	316.07
Vocational	188.07	202.93

Figure 3.3: Expected count contingency table

```
data:  tb1
X-squared = 0.40871, df = 1, p-value = 0.5226
```

Figure 3.4: Result of the Chi Square test from RStudio

The test statistic, χ^2 is obtained by using the formula:



In this case, the test statistic obtained is:

$\chi^2 = 0.40871$

Degree of freedom:

df = (2 – 1) (2 – 1) = 1

```
df = 1
```

Figure 3.5: Degree of freedom obtained from RStudio

With **α = 0.05,**

Critical value $\chi^2_{k=1,\ \alpha = 0.05} = 3.84146$

```
x2.alpha        3.84145882069413
```

Figure 3.6: Critical value obtained from RStudio

The p-value of this data set is **0.5226**.

Decision:

The test statistic is smaller than the critical value, (0.40871 < 3.84146). Which does not fall within the critical region. Hence, we fail to reject null hypothesis.

Conclusion:

There is enough evidence to show that the school accreditation and the type of school are independent.

4. Correlation Test

Test 3: Investigation of correlation analysis between parent salary per month in Indonesian Rupiah (IDR) and average grades in school.

We want to carry out a test on correlation analysis to measure the strength of the relationship between parent salary per month in Indonesian Rupiah (IDR) and average grades of the student in school. We assume that the confidence level to be 95%, significant level, $\alpha = 0.05$.

Hypothesis statement:

H0: $\rho = 0$ (no linear correlation between the parent salary per month in Indonesian Rupiah (IDR) and average grades in school.)

H1: $\rho \neq 0$ (linear correlation exists between the parent salary per month in Indonesian Rupiah (IDR) and average grades in school.)

Sample correlation coefficient formula:

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

Guide:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

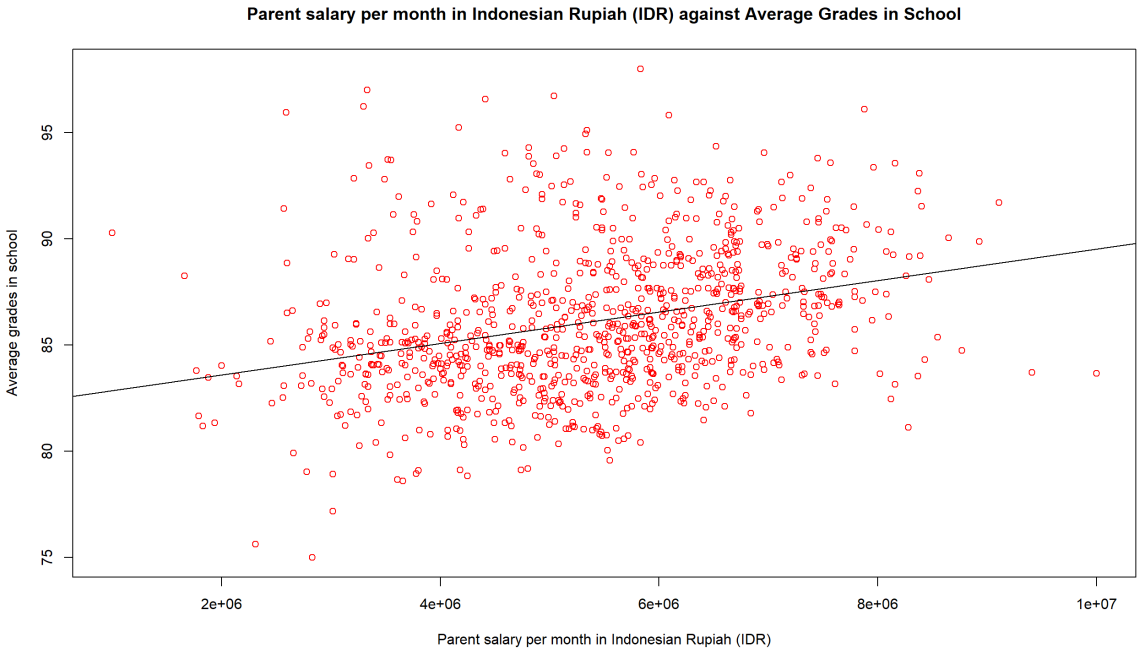


Figure 4.1: Scatter plot of Average grade in school against Parent salary per month in Indonesian Rupiah (IDR)

So, we use R studio to calculate the correlation coefficient and we find out that r = 0.3067119 which indicate weak positive linear relationship. This means that when parent salary per month increases, average grade in school also increases.

Significance test for correlation

$\alpha = 0.05$

$\alpha = 0.025$ (because it is 2 tail tests)

Correlation coefficient, r = 0.3067119

Sample size, n = 1000

Degree of freedom, df = 998

Because it is 2 tails,

Positive t-critical value: 1.9623

Negative t-critical value: -1.9623

Test statistic

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

Test statistics, t = 10.18

P-value of test statistics, t = 2.2×10^{-16}

Confidence interval of the correlation coefficient at confidence level of 95%,

CI = (0.2494621, 0.3618252)

Decision:

Since the P-value of test statistics which is 2.2×10^{-16} is less than significance level of 0.05, means that the test statistics value falls within rejection region. Hence, we reject the null hypothesis.

Conclusion:

There is sufficient evidence to conclude that there is a linear relationship between the parent salary per month in Indonesian Rupiah (IDR) and average grades in school. And because the correlation coefficient, r = 0.3067119, it shows the relationship is weak positive linear relationship.

5. Regression Test

In the regression test, we want to estimate the value the average grades in the scale of 0-100 based on the area of the parent house per square meter with a random sample of 1000 students. Hence, variables used in this test are **average grades** and **house area**. We conduct this test with a significance level of 0.05. The dependent variable, y is the average grades meanwhile the independent variable, x is the house area. The objective of this test is to test the existence of a linear relationship between y and x. Since our regression model is a linear model, hence simple linear regression is used. The changes in the values of average grades students are assumed to be caused by the changes in the values of house area.

House area(m^2) against Average grades

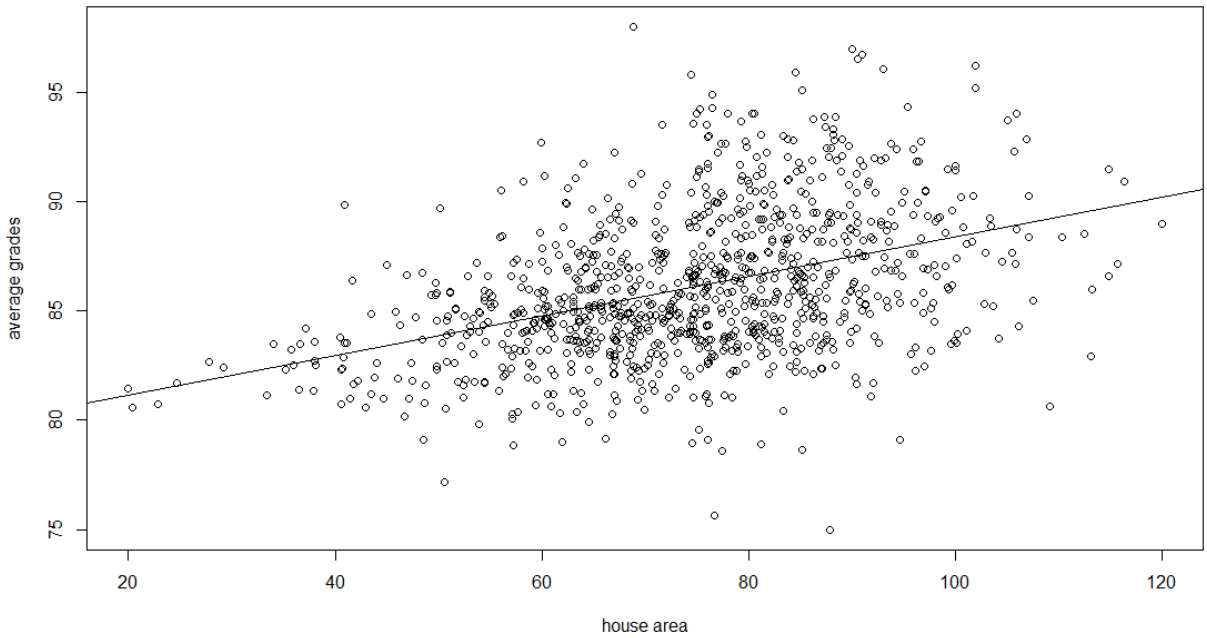


Figure 5.1: Scatter plot of House area(m^2) against Average grades

Based on Figure above, the x-axis is House area and y-axis is average grades. 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 average grades will increase as the house area increases. The equation obtained from this test is, **y = 79.3547 + 0.0905x.**

From the equation above, we can do interpretation of the intersection coefficient ß0, and slope coefficient ß1. The value of intersection coefficient:

ß0 is the estimated value of y when x = 0.

ß1 is the estimated change in the average value of y as a result of a one-unit change in x.

Based on the equation, no house had 0 per square meter, so ß0=79.3547 indicates that, for houses within the range of sizes observed,79.3547 is the portion of average grades not explained by house area. Whereas ß1=0.0905 tells us that the average value of grades increases by 0.0905 on average, for each additional one-unit house area.

From Figure 3, we can also get the coefficient of determination, R2 is 0.1677. This shows that there is only 16.77% of the variation body average grades is explained by the house area. This also shows that only some but not all the variation in average grade is explained by variation in house area.

Inference about the Slope: t-Test

Hypothesis Statement:

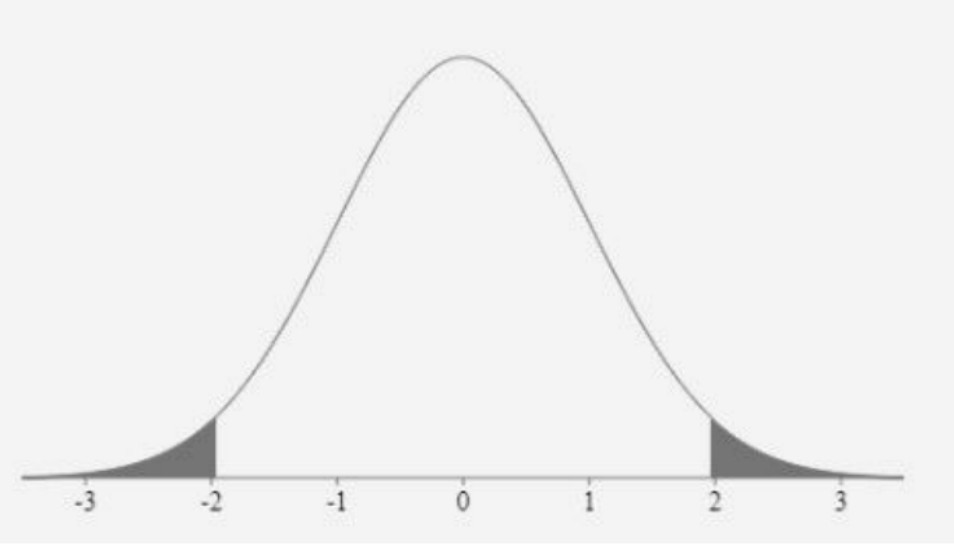
H0: $\beta 1 = 0$ (no linear regression between house area with the average grades).

H1: $\beta 1 \neq 0$ (linear regression exists between the house area with the average grades).

$\alpha = 0.05$, Sample size, n = 1000, *df* = n-2 = 998

The regression line, **y = 79.3547 + 0.0905x.**

From t-table: since this is a two-tailed test, there are two critical values:



Lower tail critical value -tα/2=0.025, df=998 = -1.9623

Upper tail critical value -tα/2=0.025, df=998 = 1.9623

Calculate test statistic by:

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

By using RStudio, we get test statistic t = 14.1827.

Decision:

Since the test statistic, t = 14.1827 is greater than tα/2=0.025, df=998 = 1.9623. It falls within the rejection region. Hence, we reject the null hypothesis.

Conclusion:

There is sufficient evidence that linear regression exists between house area and average grades of students. There is enough evidence that house area affects the student's grade.

6. Conclusion

From the one sample hypothesis testing, we could see that the students' parent monthly income salary was above the average Indonesian worker. From here we could assume that most of the students are living in a good environment and does not face any financial issues when making decision to go to the college.

From the Chi-Square Test of Independence, we have found out that the type of school and the school accreditation are independent from each other. Hence, we could assume that the type of school that the students were originated from have no relationship with its school accreditation.

From the correlation test, it shown that there is a weak relationship between parent salary per month and average grades in school. Hence, we can conclude that the parent monthly salary does not really affects their children's average grades in school.

From the regression test, we found out that there was a positive linear regression relationship exists between house area and average grades. From here, we know that house area can affect the student's grade. Hence, we can conclude that with a larger house area, it will provide a better environment for the student to study and obtain a better grade.

7. References

[1] Kaggle. (2022). *Go To College Dataset*. https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset

[2] Statistics Indonesia. (2021, June). *Average monthly net wage of employees Indonesia 2005–2021*. Statista. https://www.statista.com/statistics/1065801/indonesia-average-monthly-net-wage-of-employees/