UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143 – PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2 – GROUP 2
DATA ANALYSIS BASED ON INFERENTIAL STATISTICS

SECTION : 03 – 1 SECRH

LECTURER'S NAME : DR. Rozilawati Dollah @ Md Zain

DATE OF SUBMISSION : 3rd July 2022

No.	Name	Metric No.
1	Teo Cheen Sheng	A21EC0232
2	Haris Izudin Bin Hairul Azhar	A21EC0029
3	Lokeeshwar A/L Dinesh	A21EC3015
4	Ahmed Shukur Bin Jalaludin	A21EC0007

Table of Contents

Introduction

In this technology era, smartphone is must-have devices because they can make people's life easier. Regardless of age or sex, many people use smartphones as a means of communicating through the world wide web. Thus, about the smartphone have make many arguments on it. Some of the people will say the smartphone them to keep track of the latest thing. Other will say smartphone will cause many negative impacts. We also believe that the people nowadays already addicted on the smartphone. Therefore, smartphone technology continues to be upgraded by the current generation. Due to this, many smartphone brands have been created. But every coin has both sides, thus there is also have disadvantages of using smartphone. It will not only make the health worse, but also have negative effects on the academic results. It is common for students to become addicted to playing smartphone games until they do not care about their studies.

Research in science is the process of verifying hypotheses and applying methods to scientific activities based on real-world data. It is the process of applying scientific theory to actual events in the environment. It is usually possible to read and understand this relationship by relying on certain variables and paying attention to the evidence.

Our report is a take on the survey of the UTM student. We get 60 responders as our sample data. Our survey topic is "A Survey on the Brands of Smartphone Being Used by Students of UTM and The Effects on Their CGPA". We decided using RStudio as the tool to do the hypothesis testing. A selection of RStudio was made due to its design for data analysis and development and its ability to cover almost all the lines involved in data analysis such as its ability to perform data transformation and processing, produce excellent graphics, create interactive web-based applications, and is available in a variety of packages.

Dataset

In our project, we planned to conduct a survey on different brands of smartphone being used by UTM students and their recent CGPA. Hence, we are planning to test the data by using different types of statistical methods such as hypothesis testing two sample, correlation, regression, and chi square test (Goodness of fits test). So, from here we can conclude the data by having the different output from various method that which one is representing the best results. We have 62 dataset including boys and males who are using different type of smartphone brands and their recent CGPAs.

In hypothesis testing we focused on the difference variance among male and female students using different brands of smartphone. We used F-test to determine the difference between the two variance and we use 0.05 significance level for it. We claim that the variance of male students is greater than the variance of female students. Besides, from this test we can conclude the results from using F test.

Next, we have conducted correlation analysis using Pearson's Product-Moment Correlation method in order to measure the correlation of two variables, which are hours spent on phone per day and CGPA. Next, we also did significance test to test if there is any linear correlation between those two variables in the whole population.

Not just that, we also conducted regression analysis to obtain an equation which will be able to predict the change on the dependent variable, CGPA, if the values of independent variable, hours spent on phone per day changes. Then, we even did significance test to test whether the independent variable and dependent variable have any relationships between them.

Furthermore, we have also conducted the chi square test (One contingency unequal frequency test) to get the observed value and expected value. This is purposely for determine the difference between both value and help us to understand while interpret the two types of categorical variables which is different types of smartphones brands used by the students and the number of students.

Data Analysis

3.1 Hypothesis Testing 2-Sample

We conducted a survey on the brands of smartphone being used by the students of UTM and their recent CGPAs. There were 62 respondents for our survey where 41 of the are male students and 21 are female students. We found that the variance of male students is greater than the female students in using different type of smartphones. The variance of male student is 10.54 and the variance of female students is 4.32. To test the difference between two variance we use F- test with 0.05 significance level.

Males Students	Female Students
N1 = 41	N2 = 21
S^2 = 10.54	S^2 = 4.32

Table 1. *Number of samples and variance of female and male respondents*

Different types of smartphones	Apple	Samsung	Huawei	Vivo	Oppo	Redmi	Realme	Honor	Oneplus	Poco
Frequency	7	4	6	2	3	11	5	1	1	1
Total	**41**									

Table 2. *Data of 41 male respondents from survey*

Different types of smartphone	Apple	Samsung	Huawei	Vivo	Oppo	Redmi	Realme	Honor	Oneplus	Poco
Frequency	7	2	2	4	2	1	2	0	1	0
Total	**21**									

Table 3. *Data of 21 female respondent from survey*

Steps to Calculate the Mean and Variance value to do the F test.

I. Calculate Mean

Male Mean X' = 41/10

= 4.1

Female Mean X' = 21/10

= 2.1

II. Calculate Variance

Males			Females		
Frequency	(f – x')	$(f - x')^2$	Frequency	(f - x')	$(f - x')^2$
7	2.9	8.41	7	4.9	24.01
4	-0.1	0.01	2	-0.1	0.01
6	1.9	3.61	2	-0.1	0.01
2	-2.1	4.41	4	1.9	3.61
3	-1.1	1.21	2	-0.1	0.01
11	6.9	47.61	1	-1.1	1.21
5	0.9	0.81	2	-0.1	0.01
1	-3.1	9.61	0	-2.1	4.41
1	-3.1	9.61	1	-1.1	1.21
1	-3.1	9.61	0	-2.1	4.41
Total		94.9	Total		38.9

Table 3. *Calculation table of variance*

$$\mathbf{S2 = \sum (x - \bar{x})2 / n - 1}$$

Variance of male = 94.9 / (10 -1)

= 10.54

Variance of female = 38.9 / (10-1)

= 4.32

Significance Test for F

I. Hypothesis statement

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 > \sigma_2$

II. Significance value

= 95%

= 0.05

III. Test Statistics

We used R programming to calculate the test statistics, and we obtained the output;

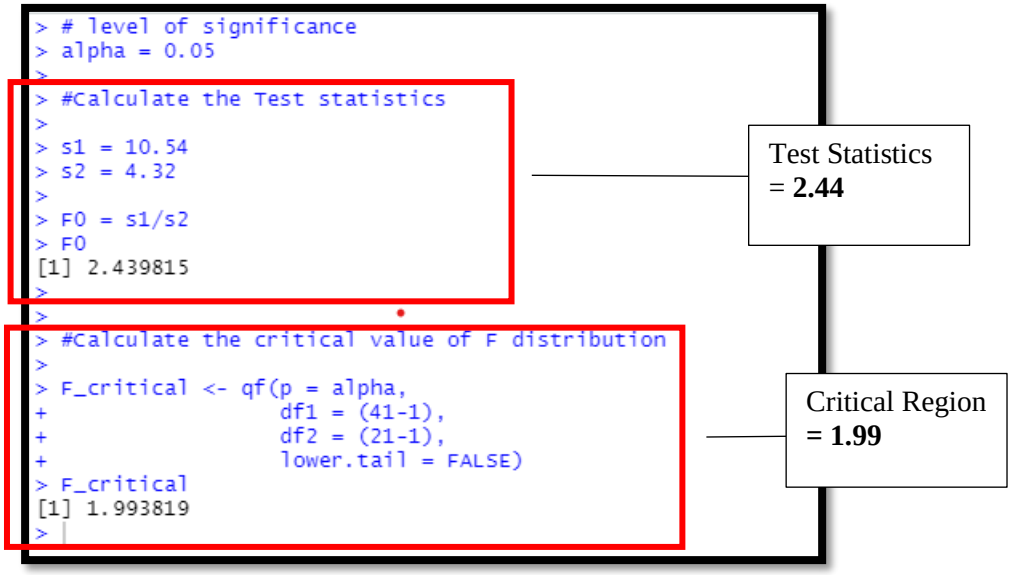

Figure 1. *Output of the test statistics and critical value calculation in R*

IV. Critical Region

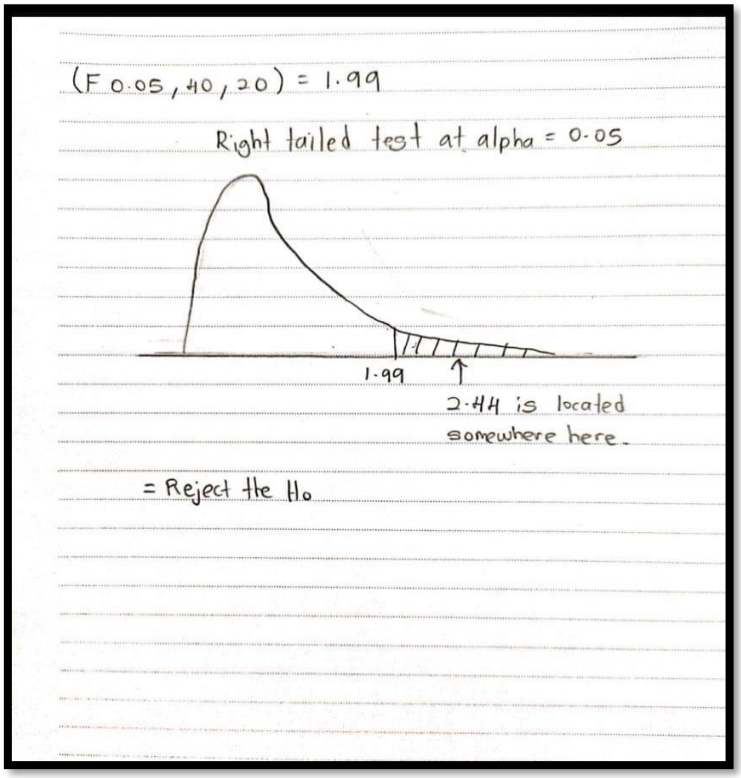

Figure 2. *F distribution graph and critical region*

V. Conclusion

Since F = 2.44 > F0.05,40,20 = 1.99, we reject the null hypothesis with 95% confidence level. Hence, we have significant evidence to conclude that variance of male students in using different brand of smartphones is larger than the variance for all female students.

3.2 Correlation

Correlation analysis functions to measure how strong is the relationship between two variables. Thus, we used this analysis to measure the strength of the relationship between the hours spent and the CGPA of the students in our sample. Since both of these variables are ratio, thus we decided to use Pearson's Product-Moment Correlation Coefficient to calculate the relationship between them, and we did significance test to measure the likelihood of the hypothesis towards the whole population.

Calculation of Pearson's Correlation Coefficient, r

I. We used R programming to calculate the Pearson Coefficient for the relationship between hours spent on phone per day with CGPA.

II. Below would be the scatter plot that we created using R programming.

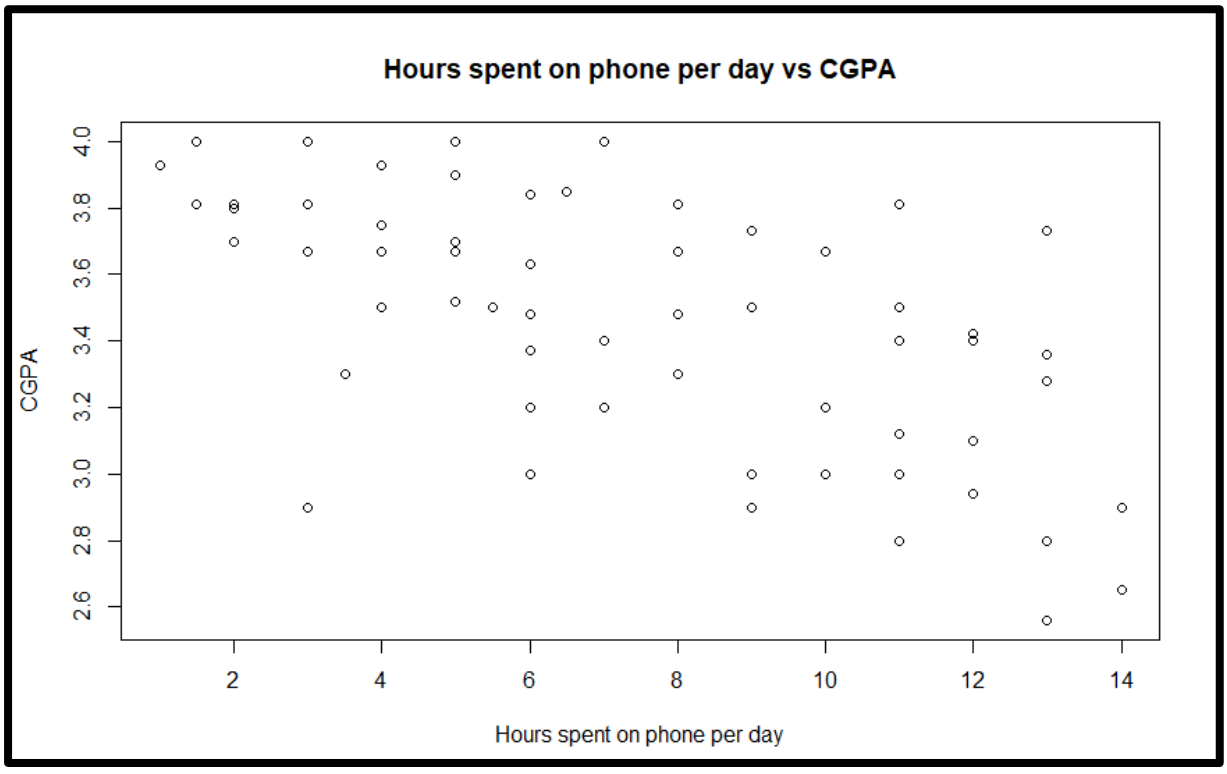

Figure 3. *Scatter plot of hours spent on phone per day and CGPA from R*

III. This is the table that we created to calculate r.

hours_spent	**CGPA**			
x	**y**	**xy**	**x^2**	**y^2**
6	3.84	23.04	36	14.7456
11	3.12	34.32	121	9.7344
13	3.36	43.68	169	11.2896
5.5	3.5	19.25	30.25	12.25
3.5	3.3	11.55	12.25	10.89
2	3.81	7.62	4	14.5161
1.5	4	6	2.25	16
4	3.5	14	16	12.25
13	3.73	48.49	169	13.9129
8	3.67	29.36	64	13.4689
11	3	33	121	9
6	3.48	20.88	36	12.1104
12	2.94	35.28	144	8.6436
11	3.4	37.4	121	11.56
9	3.73	33.57	81	13.9129
11	2.8	30.8	121	7.84
3	2.9	8.7	9	8.41
10	3.2	32	100	10.24
5	3.67	18.35	25	13.4689
13	2.56	33.28	169	6.5536
4	3.93	15.72	16	15.4449
3	3.81	11.43	9	14.5161
5	3.7	18.5	25	13.69
6	3.37	20.22	36	11.3569
6.5	3.85	25.025	42.25	14.8225
1	3.93	3.93	1	15.4449
14	2.9	40.6	196	8.41
8	3.3	26.4	64	10.89
3	3.67	11.01	9	13.4689

10	3.67	36.7	100	13.4689
7	3.2	22.4	49	10.24
12	3.4	40.8	144	11.56
7	3.4	23.8	49	11.56
11	3.81	41.91	121	14.5161
13	3.28	42.64	169	10.7584
6	3	18	36	9
5	4	20	25	16
11	3	33	121	9
9	3	27	81	9
13	2.8	36.4	169	7.84
5	3.9	19.5	25	15.21
4	3.67	14.68	16	13.4689
13	3.73	48.49	169	13.9129
2	3.7	7.4	4	13.69
7	4	28	49	16
3	4	12	9	16
1.5	3.81	5.715	2.25	14.5161
9	3.5	31.5	81	12.25
8	3.48	27.84	64	12.1104
12	3.42	41.04	144	11.6964
8	3.81	30.48	64	14.5161
11	3.5	38.5	121	12.25
2	3.8	7.6	4	14.44
6	3	18	36	9
12	3.1	37.2	144	9.61
5	3.52	17.6	25	12.3904
9	2.9	26.1	81	8.41
10	3	30	100	9
6	3.2	19.2	36	10.24
14	2.65	37.1	196	7.0225
6	3.63	21.78	36	13.1769

4	3.75	15	16	14.0625
$\sum x = 470.5$	$\sum y = 213.6$	$\sum xy = 1570.78$	$\sum x^2 = 4435.25$	$\sum y^2 = 744.7576$

Table 4. *Calculation table to obtain r value*

IV. By applying the formula of r;

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

We calculated r, by inserting our values calculated in Table (isikan siapa yg compile);

$$r = \frac{1570.78 - (470.5)(213.6)/62}{\sqrt{[(4435.25) - (470.5)^2/62][(744.7576) - (213.6)^2/62]}}$$

$$r = -0.5728$$

We also used R programming to confirm the value of r;

```
> cor(hoursCGPA$x, hoursCGPA$y)
[1] -0.5727853
```

Figure 4. *Output of calculation of r value in R*

<u>Significance Test for Pearson Correlation Obtained</u>

I. Statement of hypothesis testing:

$H_0 : \rho = 0$ ($no\ linear\ correlation\ between\ hours\ spent\ and\ CGPA$)

$H_1 : \rho \neq 0$ ($linear\ correlation\ between\ hours\ spent\ and\ CGPA\ exists$)

II. Calculation of Test Statistic, t:

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

$$r = -0.5728, \qquad n = 62$$

$$t = \frac{-0.5728}{\sqrt{\frac{1 - (-0.5728)^2}{62 - 2}}}$$

$$t = -5.4129$$

III. Critical Value (c.v.):

We choose to compare our test statistic value on the 0.05 significance level.

Degree of freedom is calculated as below;

$$df = n - 2 = 62 - 2 = 60$$

Thus, from the two-tailed t table;

$$t_{60,0.05} = 2.000$$

IV. Decision Criteria and Conclusion

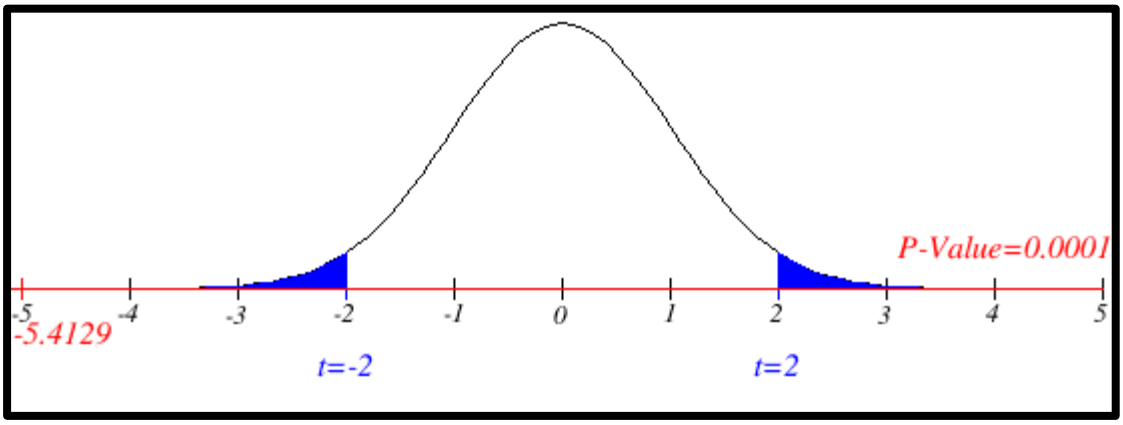

Figure 5. *t distribution graph and critical region*

Since $t = -5.4129 < t_{60,0.05} = -2.0000$, we reject the null hypothesis with significance level of 0.05. Thus, there is enough evidence to conclude that there is a linear correlation between hours spent on phone per day and CGPA at the 5% level of significance.

3.3 Regression

Regression analysis is used to predict the value of the dependent variable based on the independent variable, while also being able to explain the impact of changing the independent variable towards the dependent variable. Hence, we applied regression analysis to find the equation that will be able to predict the values of CGPA based on the hours spent on phone per day.

Estimated Regression Equation (Least Squares Equation)

I. Below would be the scatter plot with regression line that we created using R programming:

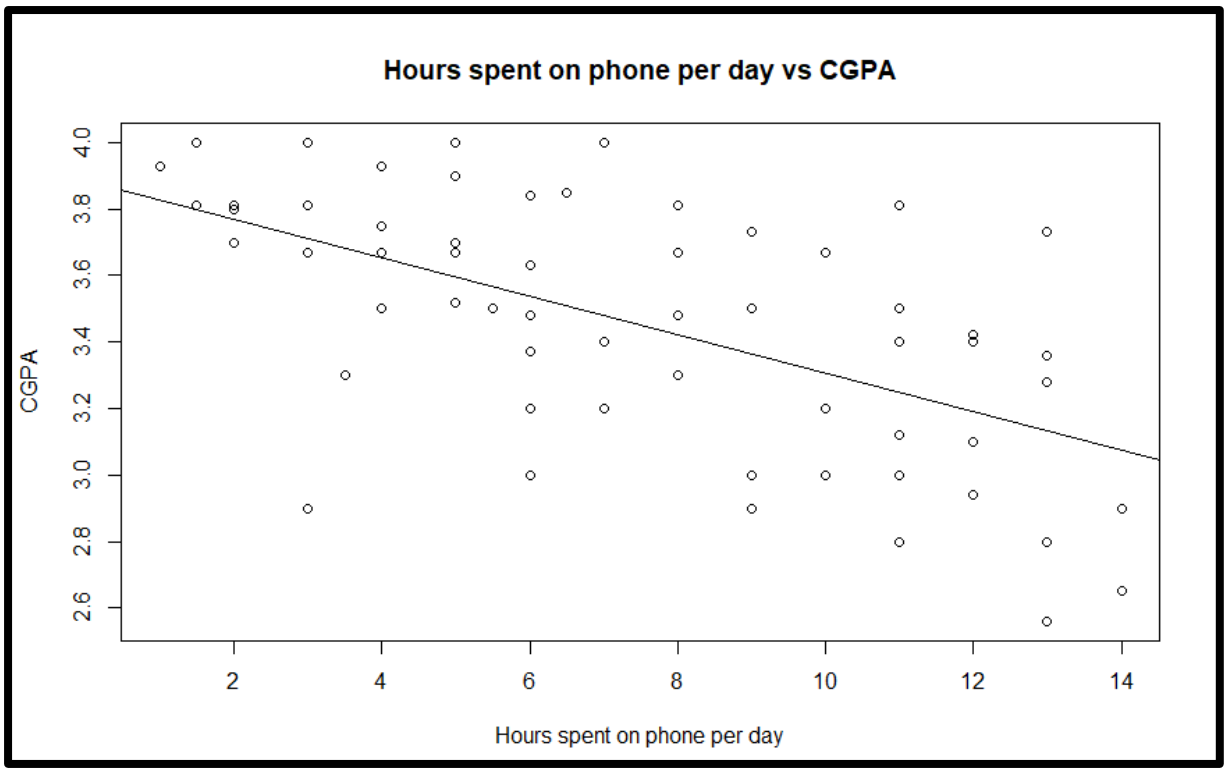

Figure 6. *Scatter plot graph of hours spent on phone per day and CGPA with regression line*

II. Calculation of b_0 and b_1:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Using the value from Table 4;

$$b_1 = \frac{1570.78 - \frac{(470.5)(213.6)}{62}}{4435.25 - \frac{470.5^2}{62}}$$

$$b_1 = -0.0580$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = \frac{213.6}{62} - (-0.0580)(\frac{470.5}{62})$$

$$b_0 = 3.8853$$

Hence, the estimated regression equation is;

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 3.8853 - 0.0580x$$

We also used R programming to confirm our b_0 and b_1 value:

```
Call:
lm(formula = y ~ x, data = hoursCGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81137 -0.22875  0.02363  0.23026  0.59877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.88541    0.09065  42.860  < 2e-16 ***
x           -0.05801    0.01072  -5.413 1.15e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3152 on 60 degrees of freedom
Multiple R-squared:  0.3281,    Adjusted R-squared:  0.3169
F-statistic:  29.3 on 1 and 60 DF,  p-value: 1.146e-06

> model

Call:
lm(formula = y ~ x, data = hoursCGPA)

Coefficients:
(Intercept)            x
    3.88541     -0.05801
```

b_0 b_1

Figure 7. *Output of b_0 and b_1 calculation in R*

Calculation of Coefficient of Determination, R^2

I. Calculate error sum of squares, SSE:

$$SSE = \sum (y - \hat{y})^2$$

By using R programming, we obtained,

$$SSE = 5.9607$$

II. Calculate regression sum of squares, SSR:

$$SSR = \sum (\hat{y} - \bar{y})^2$$

By using R programming, we obtained,

$$SSR = 2.9105$$

III. Calculate total sum of squares, SST:

$$SST = SSE + SSR$$

$$SST = 5.9607 + 2.9105$$

$$SST = 8.8712$$

IV. Calculate coefficient of determination, R^2:

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{2.9105}{8.8712}$$

$$R^2 = 0.3281$$

V. Check answer:

For single independent variable,

$$R^2 = r^2$$

$$r^2 = (-0.5728)^2$$

$$r^2 = 0.3281$$

Significance Test for Hypothesized Slope

I. Statement of hypothesis testing:

$$H_0 : \beta_1 = 0 \ (no\ linear\ relationship\ between\ hours\ spent\ \&\ CGPA)$$

$$H_1 : \beta_1 \neq 0 \ (\ linear\ relationship\ between\ hours\ spent\ \&\ CGPA\ exists)$$

II. Calculation of estimation of the standard error of the slope, s_{b_1}:

Firstly, we need to calculate the standard error of estimate, s_ε,

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

$$s_\varepsilon = \sqrt{\frac{5.9607}{62-1-1}}$$

$$s_\varepsilon = 0.3152$$

Next, we insert the value into the equation below,

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$s_{b_1} = \frac{0.3152}{\sqrt{4435.25 - \frac{470.5^2}{62}}}$$

$$s_{b_1} = 0.0107$$

III. Calculate test statistic, t:

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

$$t = \frac{-0.0580 - 0}{0.0107}$$

$$t = -5.4206$$

IV. Critical value (cv):

We chose to compare our test statistic value on the 0.05 significance level.

Degree of freedom is calculated as below;

$$df = n - 2 = 62 - 2 = 60$$

Thus, from the two-tailed t table;

$$t_{60,0.05} = 2.000$$

V. Decision Criteria and Conclusion

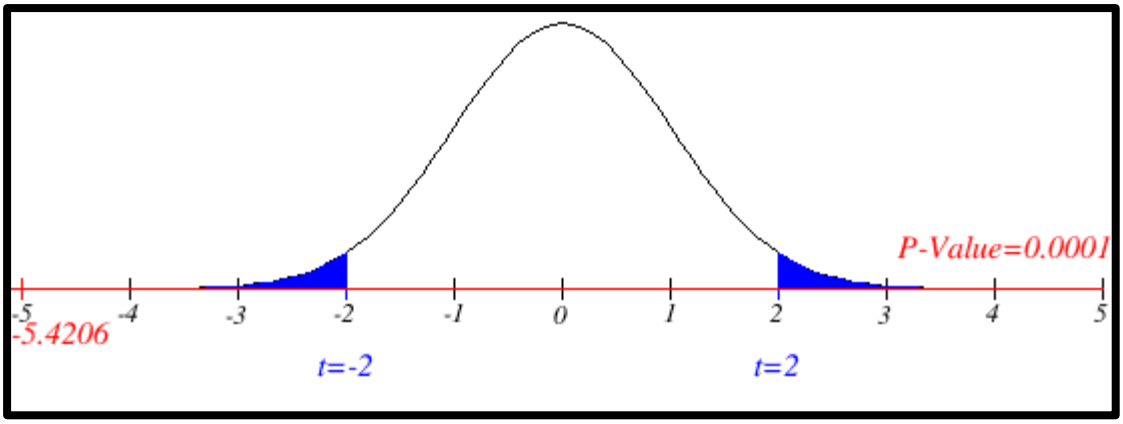

Figure 8. *t distribution graph and critical region*

Since $t = -5.4206 < t_{60,0.05} = -2.0000$, we reject the null hypothesis with significance level of 0.05. Thus, there is enough evidence to conclude that there is a relationship between hours spent on phone per day and CGPA at the 5% level of significance.

3.4 Chi-Square Test of Independence

Two of the variables that we used from the survey are smartphone brand names and the number of people who are answered this survey. Based on the survey, there are 62 people that answered the question on which smartphone brand that are they currently using. From the survey, 22% of them voted Apple, 10% voted Samsung, 13% voted Huawei, 10% voted Vivo, 8% voted Oppo, 19% voted Redmi, 11% voted Realme, 2% voted Honor, 3% voted Oneplus, and 2% voted Poco.

We determined to do a goodness of fit test. Here is the solution for this test step by step

I. Statement of hypothesis testing:

$H_0 : P_{Apple} = 0.22, P_{Samsung} = 0.10, P_{Huawei} = 0.13, P_{Vivo} = 0.10, P_{Oppo} = 0.08,$
$P_{Redmi} = 0.19, P_{Realme} = 0.11, P_{Honor} = 0.02, P_{Oneplus} = 0.03, P_{Poco} = 0.02$

H_1 : At least one of the proportions is different from the claimed value.

II. Calculate the expected frequency

	Apple	Samsung	Huawei	Vivo	Oppo	Redmi	Realme	Honor	Oneplus	Poco
O	14	6	8	6	5	12	7	1	2	1
E	13.64	6.20	8.06	6.2	4.96	11.78	6.82	1.24	1.86	1.24
$(O-E)^2$ /E	0.0095	0.0064	0.0004	0.0064	0.0003	0.0041	0.0048	0.0465	0.0105	0.0465

Table 5. *Calculation table of expected frequencies*

III. Calculate the chi-square value:

```
> ?chisq.test
> brand <- c(14,6,8,6,5,12,7,1,2,1)
> props <- c(0.22,0.1,0.13,0.1,0.08,0.19,0.11,0.02,0.03,0.02)
> chisq.test(brand, p=props)

        Chi-squared test for given probabilities

data:  brand
X-squared = 0.13547, df = 9, p-value = 1
```

Figure 9. *Calculation of chi-square value in R*

IV. Find the critical value where k=9, and α= 0.05

$X^2_{9,0.05}$= 16.919

V. State the decision

Since X^2 = 0.1354 < $X^2_{9,0.05}$= 16.919, fail to reject H_0 . There is not sufficient evidence to conclude that at least one of the proportions is different from the claimed value.

Conclusion

For conclusion we can conclude that the variance of male students in using different brand of smartphones is larger than the variance for all female students. We test it by using hypothesis testing two sample and we have the significant evidence to show that the variance of male student is larger than variance of female student.

We also do the test if there is any correlation between the hours spent and the CGPA of the students. The result of the test is having the enough evidence to conclude that there is a linear correlation between hours spent on phone per day and CGPA at the 5% level of significance.

Then, we applied regression analysis to find the equation that will be able to predict the values of CGPA based on the hours spent on phone per day. By the test, there is enough evidence to conclude that there is a relationship between hours spent on phone per day and CGPA at the 5% level of significance.

The last test that we do is a goodness of fit test to see at least one of the proportions of the smartphone brand using is different from the claimed value. By the test, there is not sufficient evidence to conclude that at least one of the proportions of the smartphone brand using is different from the claimed value.