SECI 2143-04 PROBABILITY & STATISTICAL DATA ANALYSIS – SEMESTER 2

Lecturer: Dr Aryati Bakri

Project 2

GROUP MEMBERS	MATRIC NUMBER
AFIF BIN GENARI@AZHARI	A21EC0003
MUHAMMAD ADAM HAIKAL BIN MOHD SHUHADZLEE	A21EC0062
MUHAMMAD FITRI BIN ISMAIL	A21EC0076
MUHAMMAD SYAIF ALFARIZ BIN ILYAS SUSANTO	A21EC0094

Deadline of submission: 2 July 2022

Table of contents

No	Contents	Page
1.0	Introduction	2
2.0	Dataset	3

3.0	Data Analysis
	      3.1 Hypothesis One-Sample Test
	      3.2 Correlation Test
	      3.3 Regression Test
	      3.4 Chi-Square Test of Independence
		4 - 14

4.0	Conclusion	15

5.0	Appendix
	   - Dataset
	   - Link video
	   - Link Eportfolio
		16

1.0 Introduction

In this project, we used a dataset from project 1 titled "Do video games affect students' performance?". We choose this topic because we want to investigate whether playing video games affects students' CGPA positively or negatively. Therefore, we want to prove to everyone that if the student is good at balancing their time between studying and playing video games. This video game does not affect the lesson in any way. Students can sometimes benefit from playing these video games since they can lower stress levels when a student is extremely stressed out while studying. Thus, students will feel calmer and less anxious when their emotional state is stable. So, they can continue to study comfortably and raise their CGPA further when they are not under stress. Therefore, our goal was to analyze the number of days in a week spent by students to play video games with CGPA students. With our analysis of this, we hope we can prove that by playing video games, it will not affect the CGPA of students. So we will use some methods that we have learned to analyze our study by applying what we have learned from the SECI2143 course namely hypothesis one sample test, correlation and regression tests, and chi square test of independence. We will also apply R programming and put some related graphs to further explain the analysis that we did.

2.0 Dataset

The dataset we obtained is from respondents from project 1. This dataset is used to study whether video games affect students' performance. Initially our dataset had 16 variables. So we rearranged the data and only took 5 variables to make the analysis process of our study run smoothly.

The 5 variables are the number of students, gender, age, CGPA and total days spent on games in a week. We chose these 5 variables because it is easier for us to understand to do the analysis. In addition, it is in line with what we want to analyze, which is between CGPA students and total day spent on games in a week by students. These 2 variables are important enough to determine whether video games disrupt student performance. So we will perform 4 statistical tests using Hypothesis testing 1-sample, Correlation and Regression test and Chi-Square test of independence.

3.0 Data Analysis

3.1 Hypothesis One-Sample Test

Hypothesis null, H_0 : total number of students in UTM who spend 7 days on game in a week and get a CGPA of 3.90 is 50% of all students in the UTM who spend 7 days on game in a week

Alternative Hypothesis null, H_1 : total number of students in UTM who spend 7 days on game in a week and get a CGPA of 3.90 is not equal to 50% of all students in the UTM who spend 7 days on game in a week

Total student who spend 7 days in week to play game who get cgpa of 3.90 = 10
Total student who spend 7 days in week to play game = 23

H_0 : p = 0.50
H_1 : p ≠ 0.50

α = 0.5 (two tail)

$$\hat{p} = \frac{10}{23}$$
$$= 0.4348$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{0.4348 - 0.10}{\sqrt{\frac{(0.50)(0.50)}{23}}}$$

$$= 3.21$$

Where:
$\hat{p}$ = sample proportion
p = hypothesized proportion in the null hypothesis
q = p's complement
n = sample size

P($z > 3.21$) = 1- 0.9993 = 0.0007

P ($z < -3.21$) = 0.0007

P-value = 2(0.0007) = 0.0014

Conclusion

Since 0.0014 < 0.50, reject H_0 ; there is sufficient evidence that the total number of students in UTM who spend 7 days on game in a week and get a CGPA of 3.90 is not equal to 50% of all students in the UTM who spend 7 days on game in a week.

3.2 Correlation Test

In this correlation test, we want to investigate whether there is a linear relationship between the total students who spend 7 days in a week playing games and their cgpa in the UTM. We conduct the test with a significance level of 0.05.

Correlation Between The Total Days Students Spend In a Week and Their CGPA
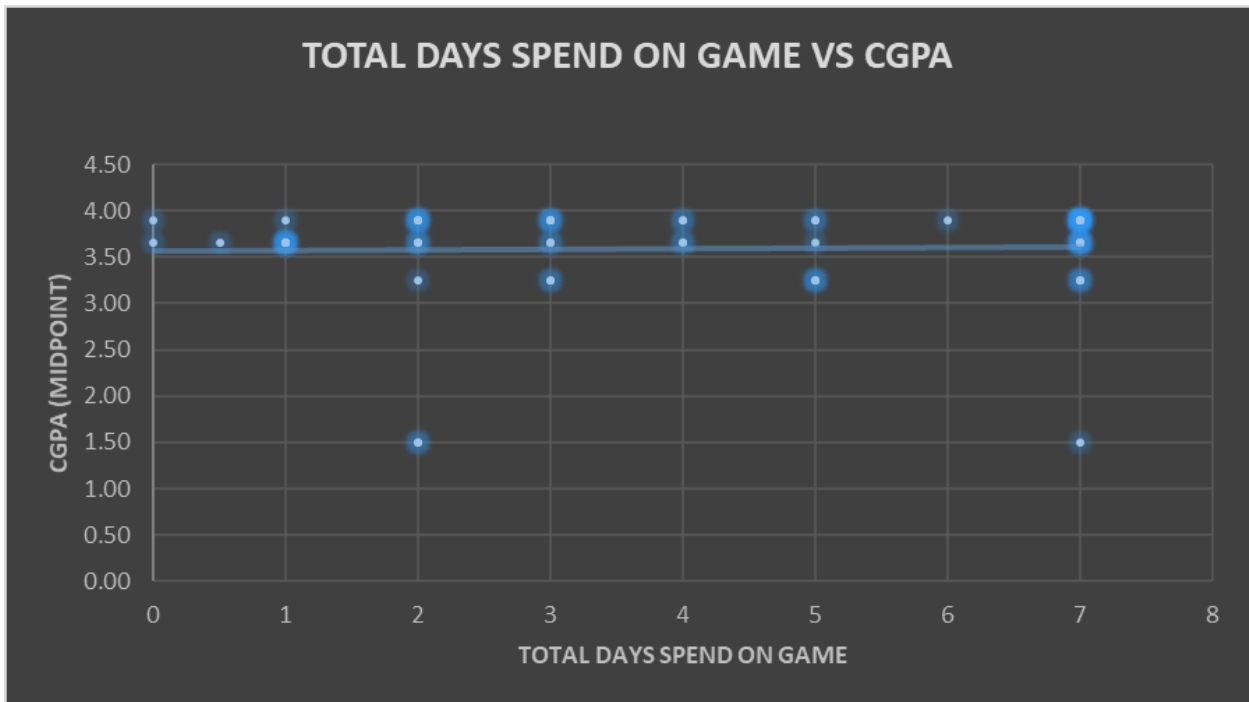
we will use the Pearson's product-moment correlation coefficient to calculate the r, which is the value to determine the strength of the linear relationship of this sample.
Calculation:

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

where:
r = sample correlation coefficient
n = sample size
x = value of independent variable
y = value of dependant variable

Based on the sample that we've been calculated in the excel file:

$$\Sigma x = 245.5 \qquad \Sigma y = 215.45 \qquad \Sigma xy = 884.075 \qquad \Sigma x^2 = 1353.25$$
$$\Sigma y^2 = 790.37$$

$$r = \frac{884.075 - 52892.975/60}{[1353.25 - (245.5)^2/60][790.37 - (215.45)^2/60]}$$

$$= 0.00043297$$

The value for r from the Pearson's product-moment correlation coefficient that we get is 0.00043297 which indicates weak positive relationship between the total days students spend on games in a week and their CGPA. After that, we proceed with the following steps.

i) State the hypothesis statement

$h_0 : \rho = 0$ (no linear correlation)

$h_1 : \rho \neq 0$ (there is linear correlation)

ii) Find the critical value

$\alpha = 0.05 \qquad$ degree of freedom($d.f$) = n - 2 = 60 - 2 = 58

$t_{0.05, 58} = 1.645$

iii) Calculate the test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.00043297}{\sqrt{\frac{1-(0.00043297)^2}{60-2}}} = 0.0032974$$

iv) Decision: Since $t_{test\ statistic} = 0.0032974 < t_{critical\ value} = 1.645$, we fail to reject h_0.

v) Conclusion: There is no sufficient evidence to support that there is a linear relationship between the total days students spend on games in a week and their CGPA at the 5% level of significance.

3.3 Regression Test

      The simple regression test we conducted is to determine if the total day spent on a game in a week could be used to predict the student's CGPA. As such, we set the total day spent on a game in a week as an independent variable (x-axis) and student's CGPA as a dependent variable(y-axis). As usual, we set the confidence level as $\alpha = 0.05$. This test is partly used to detect if a linear relationship is present between the two aforementioned variables.

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

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

Based on the formula above b0 is the estimated average value of y when the value of x is zero and b1 is the estimated change in the average value of y as a result of a one-unit change in x.

$\Sigma x = 245.5$ $\Sigma y = 215.45$ $\Sigma xy = 884.075$ $\Sigma x^2 = 1353.25$

$$b1 = \frac{(884.075) - \frac{(245.5)(215.45)}{60}}{(1353.25) - \frac{(245.5)^2}{60}}$$

$$= 0.007241424$$

$\bar{y} = 3.59083$ $\bar{x} = 4.09167$

$$b0 = 3.59083 - (0.007241424)(4.09167)$$
$$= 3.56$$

From the calculation above we can obtained the estimated regression equation:

$\hat{y} = 0.0072x + 3.56$

Calculation for Coefficient of Determination, R^2

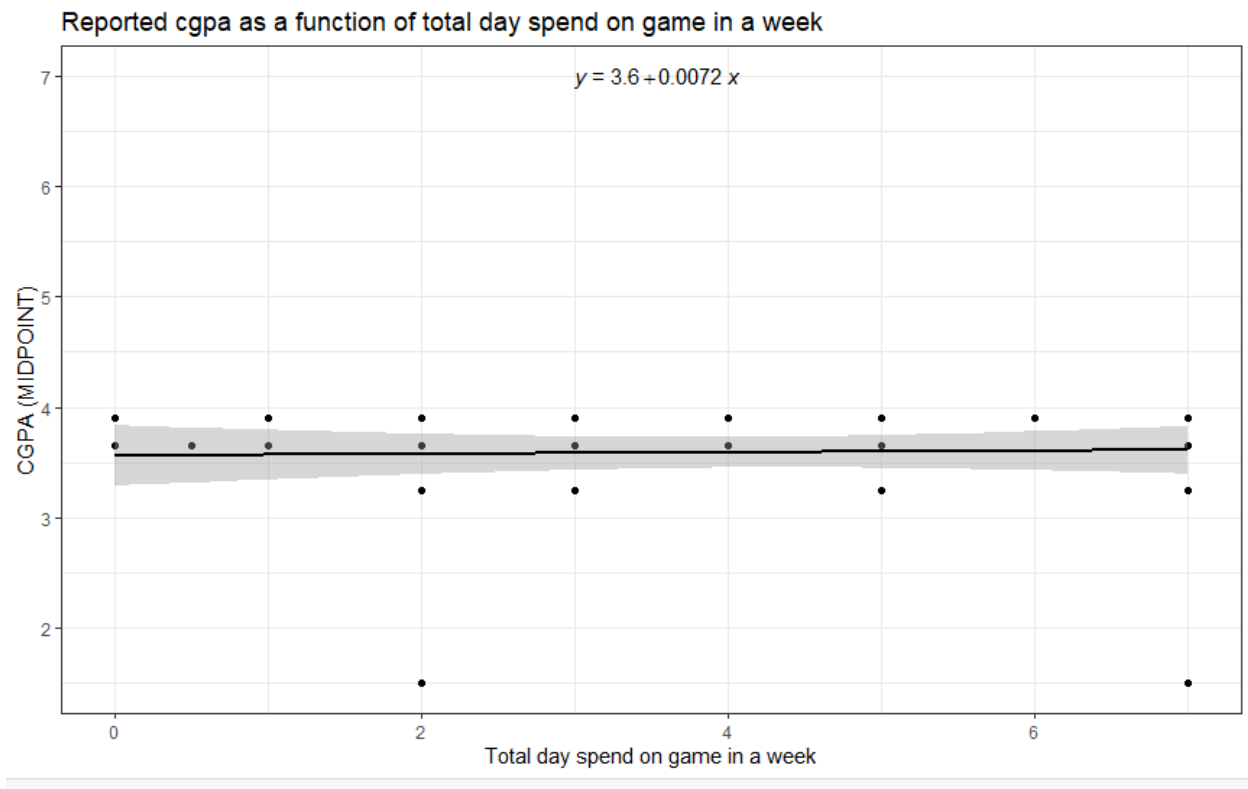$$R^2 = \frac{SSR}{SST} = \frac{sum\ of\ squares\ explained\ by\ regression}{total\ sum\ of\ squares}$$

As for SSR and SST we have calculated it using excel which the value is
$SSR = 0.018192$
$SST = 16.722458$

$$R^2 = \frac{0.018192}{16.722458} = 0.0011$$

Regression between total spent on a game in a week and student's CGPA using excel

Reported cgpa as a function of total day spend on game in a week

$y = 3.6 + 0.0072\,x$

Based on the value of $R^2 = 0.0011$ we can say that the relationship between total day spent on a game in a week(independent variable) and student's CGPA in midpoint (dependent variable) is a weak relationship since the value is closer to 0. Apart from that we can say that only 0.11% variation in student's CGPA in midpoint is explained by variation in total spent on a game in a week.Next, as for value of b1 which is the slope coefficient b1=0.0072 tells us that the average

value for student's CGPA is increase by 0.0072 for each additional day in total day spent on a game in a week or we can say that when the students spent more days in playing games their CGPA will increase by 0.0072. As for the value of b0 which is the value of y-intercept b0=3.56 tells us that if the student spent 0 days playing video games or we can say that the students did not play any games the expected CGPA that they probably obtain is 3.56 .

Standard Error of Estimate :

SSE=16.70428448

$$S\epsilon = \sqrt{\frac{SSE}{n-k-1}}$$

$$= \sqrt{\frac{16.70428448}{60}} = 0.52764$$

Standard Deviation of the Regression Slope:

$$S_{b1} = \frac{S\epsilon}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

$$= \frac{0.52764}{\sqrt{1353.25 - \frac{(245.5)^2}{60}}} = 0.03$$

Inference about the slope t test:

$H_0: \beta 1 = 0$ (*no linear relationship*)
$H_1: \beta 1 \neq 0$ (*linear relationship does exist*)

Test statistic:
$$t = \frac{b_1 - \beta 1}{S_{b1}}$$
$$= \frac{0.0072 - 0}{0.03} = 0.24$$
d.f= 60-2 = 58
$\alpha = 0.05$
$t_{0.05/2} = 2.0017$ (critical value)

Decision:
Since test statistic $<$ critical value, thus we fail to reject H_0 at $\alpha = 0.05$

Conclusion:
There is sufficient evidence that the total day spent on a game in a week is independent of the student's cgpa.

3.4 Chi-Square Test of Independence

a) **Hypothesis Statement**

H_0 : The total day spent on a game in a week is independent of the student's cgpa.

H_1 : The total day spent on a game in a week is dependent on the student's cgpa.

b) **Critical Value**

Test Statistic :

$$x^2 = \sum \frac{(o_i - e_i)^2}{e_i},$$

o = Observed Count
e = Expected Count

Degree of Freedom, df = (r - 1) (c - 1),

Critical Value, α = 0.05
Degree of Freedom, df = (10 - 1)(2 - 1) = 9

c) Calculate Expected Count and Test Statistic

Expected Count calculation formula:

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

Total day spend on game in a week	CGPA(MIDPOINT)	Total
4	3.90	7.9
7	3.25	10.25
2	3.65	5.65
7	3.65	10.65
5	3.25	8.25
2	3.90	5.9
0.5	3.65	4.15
1	3.90	4.9
2	3.90	5.9
7	3.65	10.65
37.5	36.70	74.2

Sample Data

Cell, ij	Observed Count, Oij	Expected Count, Eij	(O-E)^2 / E
1,1	4	3.9926	1.37154E-05
1,2	3.90	3.9074	1.40144E-05
2,1	7	5.1803	0.639211646
2,2	3.25	5.0697	0.653156615
3,1	2	2.8555	0.256305463
3,2	3.65	2.7945	0.26190025
4,1	7	5.3824	0.486145541
4,2	3.65	5.2676	0.496740405
5,1	5	4.1695	0.165422773
5,2	3.25	4.0805	0.169030817
6,1	2	2.9818	0.323271594
6,2	3.90	2.9182	0.330317058
7,1	0.5	2.0974	1.216595194
7,2	3.65	2.0526	1.243148573
8,1	1	2.4764	0.880211985
8,2	3.90	2.4236	0.899388084
9,1	2	2.9818	0.323271594
9,2	3.90	2.9182	0.330317058
10,1	7	5.3824	0.486145541
10,2	3.65	5.2676	0.496740405
		x^2 =	9.657348328

Table to calculated Expected count and Test Statistic

d) State the decision

$$\text{Test Statistic, } x^2 = 9.66$$
$$\text{Critical value, } x^2_{k=9, \alpha=0.05} = 16.919$$

Decision:
Since, test statistic value < critical value, thus we fail to reject H_0 at $\alpha = 0.05$.

Conclusion:
There is sufficient evidence that the total day spend on game in a week is independent of the student's cgpa.

4.0 Conclusion

In conclusion, we have used various techniques that have been learned in this course to calculate and analyze the data we selected, that is the analysis of whether video games affect student performance. The dataset we obtained is from the respondent of project 1. So for this project 2 we have done pre-processing of data by taking 5 out of 16 variables only. We do so to simplify the calculation process while doing statistical tests.

Based on the four statistical tests we performed, we can conclude that there is sufficient evidence that the total number of students in UTM who spend 7 days on game in a week and get a CGPA of 3.90 is not equal to 50% of all students in the UTM who spend 7 days on game in a week. Beside that, there is no sufficient evidence to support that there is a linear relationship between the total days students spend on games in a week and their CGPA at the 5% level of significance. Next, There is sufficient evidence that the total day spent on a game in a week is independent of the student's cgpa.

We would like to thank Dr Aryati for guiding us in completing this project 2. We have been able to get a lot of output while doing this task. Among them is that we can apply R programming, how to read data, how to do data processing and can further strengthen skills in performing data analysis.

5.0 Appendix

Dataset:
https://docs.google.com/spreadsheets/d/1pXm2cseC_LrOEppVmD8fY_eDo_8CDhQutpOCbvoAu-k/edit?usp=sharing

Link video: https://www.youtube.com/watch?v=4H6x0I62uj0

Link eportfolio:
Afif   - https://eportfolio.utm.my/user/afif-bin-genari-azhari
Adam - https://eportfolio.utm.my/user/muhammad-adam-haikal-bin-mohd-
Fitri   - https://eportfolio.utm.my/user/muhammad-fitri-bin-ismail
Syaif - https://eportfolio.utm.my/view/view.php?t=v79UWIpt5gOAq1LohVxu