UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

BUM2413 APPLIED STATISTICS
SEM I 2024/2025

ASSIGNMENT COVER

SECTION: 12P GROUP **NAME**: OUTLIERS

GROUP MEMBERS (State the PIC)	**STUDENT ID**	**SIGNATURE**
1. HAZIRAH BINTI ERMON CHATIB (PIC)	CB23039	
2. NOORA HANIM BINTI AZMI	CB23040	
3. SITI NAJWA BINTI BAKRI	CB23013	
4. UMI MAISARAH BINTI MOHD AFENDI	CB23036	
5. WARDAH WAFIN BINTI YACOB	CB23076	
LECTURER: Ms. ZAHIRAH BINTI MD NOOR		
SUBMISSION DATE: 6 JUNE 2025		

FOR EXAMINER USE ONLY					
Question	**Marks**	**Your Marks**	**Question**	**Marks**	**Your Marks**
1	**2**		**7**	**8**	
2	**1**		**8**	**2**	
3	**2**		**9**	**2**	
4	**7**		**10**	**4**	
5	**3**		**11**	**24**	
6	**4**		**12**	**1**	
			TOTAL	**60**	

Table of Contents

1.0 INTRODUCTION 3
2.0 DATA DESCRIPTION 3
3.0 DATA COLLECTION METHOD 3
4.0 DATA SUMMARY METHOD 6
5.0 DATA NORMALITY 9
6.0 HYPOTHESIS TESTING FOR ONE POPULTAION MEAN 14
7.0 HYPOTHESIS TESTING FOR TWO POPULATION MEANS 15
8.0 CONCLUSION 18
9.0 REFERENCES 19
APPENDICES 20

1.0 INTRODUCTION

Identify a topic/ problem that you are interested in studying. Provide a brief description of your study and state at least ONE (1) study objective. (2 Marks)

Students' daily routines have become increasingly reliant on digital devices such as smartphones, tablets, and laptops. While these gadgets are necessary for learning and communication, excessive screen time can harm students' academic performance because of distractions, decreased study time, and poor sleep quality.

This project seeks to investigate the relationship between students' time spent on study and their academic performance. The project will investigate whether there is a significant link between students' study time and academic success by collecting and evaluating data on students' daily study time and academic performance which is their CGPA. The findings are expected to provide useful insights into how screen time habits can affect educational outcomes.

2.0 DATA DESCRIPTION

State the population of the study. (1 Mark)

The population is all students in University Malaysia Pahang Al-Sultan Abdullah (UMPSA).

Determine a single quantitative variable that is related to your chosen problem. Identify the type of level of measurement for the variable. (2 Marks)

The single quantitative variable chosen is between students' time spent on study.

The type of measurement for this variable is the ratio level.

3.0 DATA COLLECTION METHOD

Divide the data collected into two significant groups (e.g.: gender (male/female), faculty, year of study, etc.) that are related to the study. The sample size is at least 50 observations for each group.

(i) State the name of the groups. (1 Mark)

Two significant groups of students will be categorized according to students' daily study time which is daily study time less than 2 hours and daily study time more than 2 hours.

(ii) Present the data collected according to the groups in a table. (2 marks)

GROUP 1		GROUP 2	
daily study time less than 2 hours	**CGPA**	**daily study time more than 2 hours**	**CGPA**
1	2.49	**1**	3.47
2	3.60	**2**	3.87
3	3.50	**3**	3.78
4	3.50	**4**	3.10
5	3.50	**5**	3.50
6	3.22	**6**	4.00
7	3.00	**7**	3.60
8	3.30	**8**	3.50
9	3.30	**9**	3.66
10	3.20	**10**	4.00
11	3.70	**11**	3.40
12	3.70	**12**	3.80
13	3.23	**13**	3.12
14	2.00	**14**	3.40
15	3.60	**15**	3.40
16	3.57	**16**	3.49
17	3.80	**17**	3.50
18	3.57	**18**	3.60
19	3.30	**19**	3.65
20	3.00	**20**	3.80
21	3.30	**21**	3.80
22	3.30	**22**	3.80
23	3.40	**23**	4.00
24	3.57	**24**	3.00
25	3.20	**25**	3.30
26	3.50	**26**	3.30

27	3.00	**27**	3.80
28	3.33	**28**	3.80
29	2.95	**29**	3.94
30	3.00	**30**	3.20
31	3.10	**31**	3.80
32	3.20	**32**	2.90
33	3.50	**33**	3.30
34	3.00	**34**	3.50
35	3.46	**35**	3.20
36	3.00	**36**	3.33
37	3.77	**37**	3.40
38	3.80	**38**	3.63
39	3.00	**39**	3.67
40	3.00	**40**	3.90
41	3.00	**41**	3.40
42	3.20	**42**	3.60
43	3.50	**43**	3.60
44	3.65	**44**	3.70
45	3.33	**45**	3.70
46	3.50	**46**	3.23
47	3.82	**47**	3.30
48	3.60	**48**	3.75
49	3.70	**49**	3.00
50	2.05	**50**	3.50

(iii) Identify the method of data collection being used. Provide significant evidence. (2 Marks)

The data collection method is a online survey form. The Google Form link provided: https://forms.gle/u4hvCbX1bbUHso8r5 serves as strong evidence of this method. This form allows respondents to answer questions about their daily study time and CGPA, making it an efficient method for getting data from a large student population.

(iv) State the sampling method you use to collect the data. Explain the sampling method process. (2 Marks)

The sampling method used is voluntary sampling. The online survey link is distributed to all University Malaysia Pahang Al-Sultan Abdullah (UMPSA) students via WhatsApp, group chats and online post. This method allowed students who were interested in the research topic to voluntarily complete the survey, resulting in a diverse set of data on students' daily study time and its potential link to students' CGPA.

4.0 DATA SUMMARY METHOD

For each set of data, obtain the descriptive statistics using Microsoft Excel. Then, summarise the measures of central tendency (mean, median, mode, midrange) and measures of variation (range, standard deviation, variance) in a table. (3 Marks)

CGPA for daily study time less than 2 hours	
Mean	3.2962
Standard Error	0.054703572
Median	3.315
Mode	3
Standard Deviation	0.386812669
Sample Variance	0.149624041
Kurtosis	3.141968267
Skewness	-1.456820799
Range	1.82
Minimum	2
Maximum	3.82
Sum	164.81
Count	50
Largest(1)	3.82
Smallest(1)	2
Confidence Level(95.0%)	0.109930944

CGPA for daily study time more than 2 hours	
Mean	3.5398
Standard Error	0.039400083
Median	3.55
Mode	3.8
Standard Deviation	0.278600658
Sample Variance	0.077618327
Kurtosis	-0.561538379
Skewness	-0.316320839
Range	1.1
Minimum	2.9
Maximum	4
Sum	176.99
Count	50
Largest(1)	4
Smallest(1)	2.9
Confidence Level(95.0%)	0.079177431

Group name	**Measures of central tendency**		**Measures of variation**	
Group daily study time less than 2 hours	**Mean,** $\overline{x_1}$	3.2962	**Range**	1.8200
	Median	3.3150	**Standard deviation,** s_1	0.3868

	Mode	3.0000	**Variance**	0.1496
	Midrange	2.9100		
Group daily study time more than 2 hours	**Mean,** $\overline{x}_2$	3.5398	**Range**	1.1000
	Median	3.5500	**Standard deviation,** s_2	0.2786
	Mode	3.800	**Variance**	0.0776
	Midrange	3.4500		

Compare and comment on the measures of central tendency and measures of variation between Group 1 and Group 2. (4 Marks)

$\overline{x}_2 > \overline{x}_1$, daily study time more than 2 hours have higher CGPA than daily study time less than 2 hours.

$s_1 > s_2$ CGPA is more consistent for daily study time less than 2 hours than daily study time more than 2 hours.

Construct box plots for the two sets of data on the same axis. Identify the shape of the distribution for each boxplot. Compare and comment on the average and variability of the boxplots. (8 Marks)

Summary	Less than 2 Hours	More than 2 Hours
Min	2.000	2.900
Q1	3.025	3.3475
Median (Q2)	3.315	3.550
Q3	3.570	3.795
Max	3.820	4.00
Mean	3.2962	3.5398
IQR	0.545	0.4475
Outlier	No Outlier	No Outlier

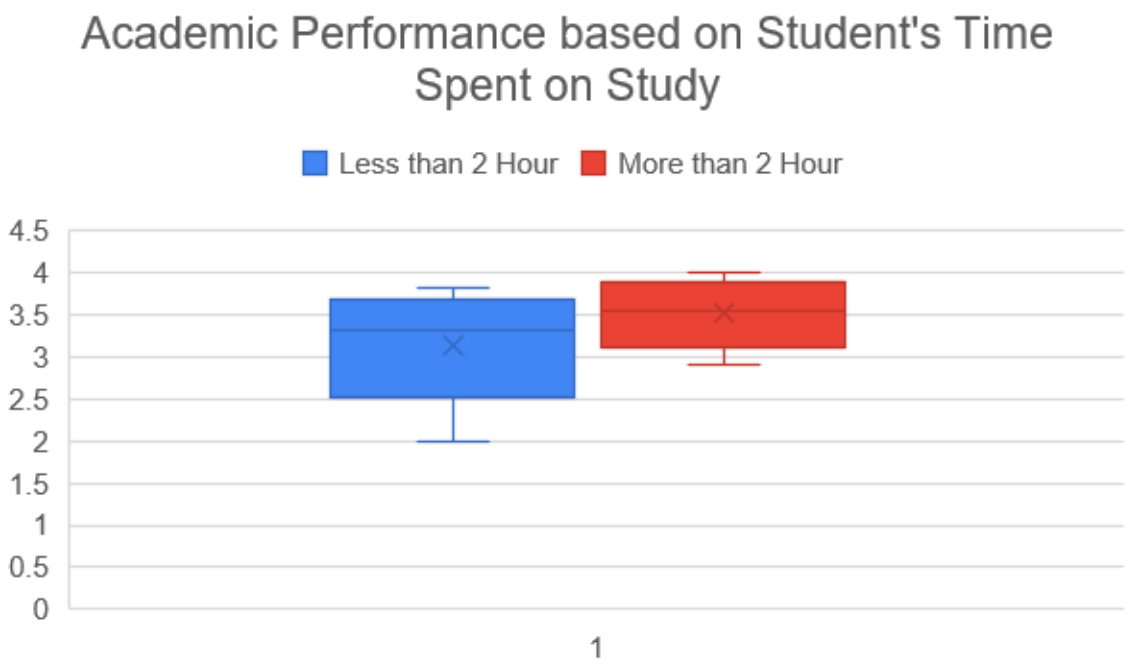


Figure: Box Plot of Academic Performance based on Student's Time Spent on Study

Shape of Distribution

Less than 2 hours: Left-Skewed

More than 2 hours: Left-Skewed

Mean/Average

- The average of academic performance for student who spent More than 2 Hours is higher which is 3.5498 than student who spent Less than 2 Hours which is 3.2962.
- On average, students who spent More than 2 Hours has higher academic performance compared to students who spent Less than 2 Hours.

Variability

- Based on IQR value for Less than 2 Hours = 0.545, meaning the middle 50% of the data is spread over a range of 0.545 units. Meanwhile, IQR value for More than 2 Hours = 0.4475, which means that the middle 50% of the data points are more tightly clustered.
- The group studying more than 2 hours shows **slightly less variability** in academic performance (lower IQR of **0.4475**) compared to the group studying less than 2 hours (IQR of **0.545**).

What is the best measure of central tendency to describe your data? Give a reason. (2 Marks)

The best measure of central tendency is median. This because median is less affected by extreme value or outliers.

What is the best measure of variation to describe your data? Give a reason. (2 Marks)

The best measure of variation is Interquartile Range IQR. This is because IQR is less influence or not easily affected by extreme value or outliers.

5.0 DATA NORMALITY

Construct a normal probability plot for each data set. Do the data appear to come from an approximately normal distribution? (4 Marks)

NO.	Daily study time less than 2 hours		Daily study time more than 2 hours	
	Original data	Sorted data	Original data	Sorted data
1	2.49	2.00	3.47	2.90
2	3.60	2.05	3.87	3.00
3	3.50	2.49	3.78	3.00
4	3.50	2.95	3.10	3.10
5	3.50	3.00	3.50	3.12
6	3.22	3.00	4.00	3.20
7	3.00	3.00	3.60	3.20
8	3.30	3.00	3.50	3.23
9	3.30	3.00	3.66	3.30
10	3.20	3.00	4.00	3.30
11	3.70	3.00	3.40	3.30
12	3.70	3.00	3.80	3.30
13	3.23	3.00	3.12	3.33
14	2.00	3.10	3.40	3.40
15	3.60	3.20	3.40	3.40
16	3.57	3.20	3.49	3.40

17	3.80	3.20	3.50	3.40
18	3.57	3.20	3.60	3.40
19	3.30	3.22	3.65	3.47
20	3.00	3.23	3.80	3.49
21	3.30	3.30	3.80	3.50
22	3.30	3.30	3.80	3.50
23	3.40	3.30	4.00	3.50
24	3.57	3.30	3.00	3.5
25	3.20	3.30	3.30	3.50
26	3.50	3.33	3.30	3.60
27	3.00	3.33	3.80	3.60
28	3.33	3.40	3.80	3.60
29	2.95	3.46	3.94	3.60
30	3.00	3.50	3.20	3.63
31	3.10	3.50	3.80	3.65
32	3.20	3.50	2.90	3.66
33	3.50	3.50	3.30	3.67
34	3.00	3.50	3.50	3.70
35	3.46	3.50	3.20	3.70
36	3.00	3.50	3.33	3.75
37	3.77	3.57	3.40	3.78
38	3.80	3.57	3.63	3.80
39	3.00	3.57	3.67	3.80
40	3.00	3.60	3.90	3.80
41	3.00	3.60	3.40	3.80
42	3.20	3.60	3.60	3.80
43	3.50	3.65	3.60	3.80
44	3.65	3.70	3.70	3.80
45	3.33	3.70	3.70	3.87
46	3.50	3.70	3.23	3.90

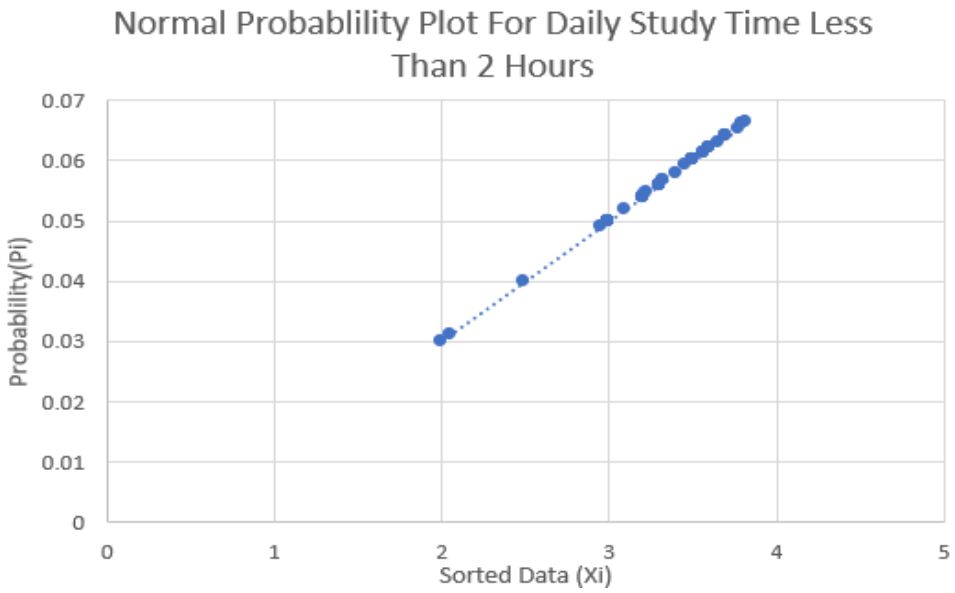
47	3.82	3.77	3.30	3.94
48	3.60	3.80	3.75	4.00
49	3.70	3.80	3.00	4.00
50	2.05	3.82	3.50	4.00

Daily study time less than 2 hours			**Daily study time more than 2 hours**		
i	**Sorted data of CGPA, X_i**	$pi = \frac{(i-0.5)}{n}$	i	**Sorted data of CGPA, X_i**	$pi = \frac{(i-0.5)}{n}$
1	2.00	0.0300	**1**	2.90	0.0480
2	2.05	0.0310	**2**	3.00	0.0500
3	2.49	0.0398	**3**	3.00	0.0500
4	2.95	0.0490	**4**	3.10	0.0520
5	3.00	0.0500	**5**	3.12	0.0524
6	3.00	0.0500	**6**	3.20	0.0540
7	3.00	0.0500	**7**	3.20	0.0540
8	3.00	0.0500	**8**	3.23	0.0546
9	3.00	0.0500	**9**	3.30	0.0560
10	3.00	0.0500	**10**	3.30	0.0560
11	3.00	0.0500	**11**	3.30	0.0560
12	3.00	0.0500	**12**	3.30	0.0560
13	3.00	0.0500	**13**	3.33	0.0566
14	3.10	0.0520	**14**	3.40	0.0580
15	3.20	0.0540	**15**	3.40	0.0580
16	3.20	0.0540	**16**	3.40	0.0580
17	3.20	0.0540	**17**	3.40	0.0580
18	3.20	0.0540	**18**	3.40	0.0580
19	3.22	0.0544	**19**	3.47	0.0594
20	3.23	0.0546	**20**	3.49	0.0598

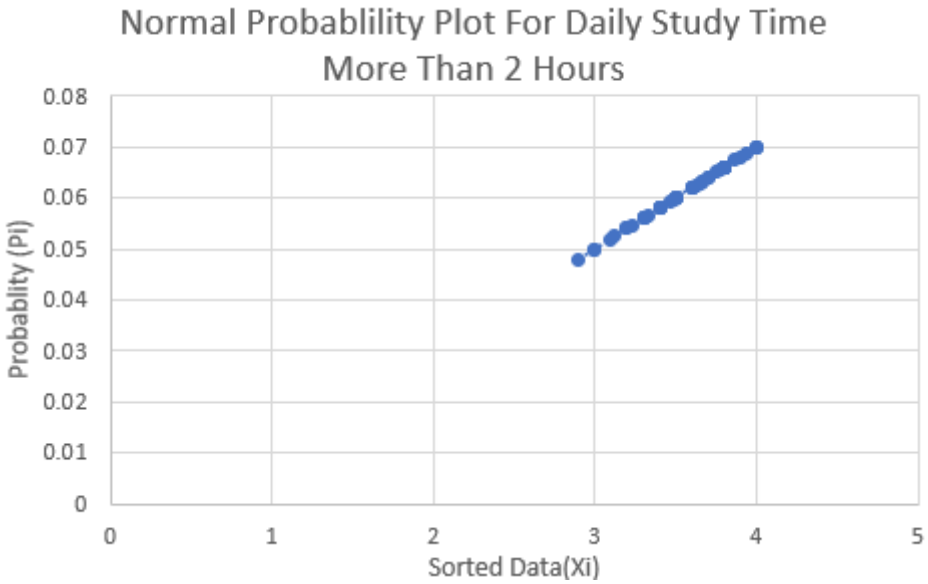
21	3.30	0.0560	**21**	3.50	0.0600
22	3.30	0.0560	**22**	3.50	0.0600
23	3.30	0.0560	**23**	3.50	0.0600
24	3.30	0.0560	**24**	3.5	0.0600
25	3.30	0.0560	**25**	3.50	0.0600
26	3.33	0.0566	**26**	3.60	0.0620
27	3.33	0.0566	**27**	3.60	0.0620
28	3.40	0.0580	**28**	3.60	0.0620
29	3.46	0.0592	**29**	3.60	0.0620
30	3.50	0.0600	**30**	3.63	0.0626
31	3.50	0.0600	**31**	3.65	0.0630
32	3.50	0.0600	**32**	3.66	0.0632
33	3.50	0.0600	**33**	3.67	0.0634
34	3.50	0.0600	**34**	3.70	0.0640
35	3.50	0.0600	**35**	3.70	0.0640
36	3.50	0.0600	**36**	3.75	0.0650
37	3.57	0.0614	**37**	3.78	0.0656
38	3.57	0.0614	**38**	3.80	0.0660
39	3.57	0.0614	**39**	3.80	0.0660
40	3.60	0.0620	**40**	3.80	0.0660
41	3.60	0.0620	**41**	3.80	0.0660
42	3.60	0.0620	**42**	3.80	0.0660
43	3.65	0.0630	**43**	3.80	0.0660
44	3.70	0.0640	**44**	3.80	0.0660
45	3.70	0.0640	**45**	3.87	0.0674
46	3.70	0.0640	**46**	3.90	0.0680
47	3.77	0.0654	**47**	3.94	0.0688
48	3.80	0.0660	**48**	4.00	0.0700
49	3.80	0.0660	**49**	4.00	0.0700

50	3.82	0.0664	**50**	4.00	0.0700

Normal Probability plot for daily study time less than 2 hours



Normal Probability plot for daily study time more than 2 hours



Comment:

Both the Group 1 (daily study time less than 2) and Group 2 (daily study time) are normal probability plot since the data lie approximately on a straight line. The data are normally distributed population.

6.0 HYPOTHESIS TESTING FOR ONE POPULTAION MEAN

Hypothesis testing is one of the inferential statistics in statistical analysis. The parameters are the population mean, proportion, variance and standard deviation. Assuming that the data obtained in (4) is a normally distributed population, answer the following questions using the P-value approach and Microsoft Excel.

a. Create a situation and conduct hypothesis testing for one population mean from one of the groups. (8 Marks)

Group Less than 2 hours is chosen for hypothesis testing. Since the total population mean for the CGPA of UMPSA students is not available, it is assumed that the CGPA of UMPSA students is at least 3.30. To test this hypothesis, a sample of 50 UMPSA students that study less than 2 hours a day was collected, and the calculated mean is 3.2962. A 0.05 level of significance is used to test the hypothesis.

Step 1: $H_0 : \mu \geq 3.30$ (Claim)
$H_1 : \mu < 3.30$

Step 2: σ^2 is unknown, n ≥ 30, z test, left-tailed test

CGPA of UMPSA Students That Study Less than 2 Hours	
Mean	3.2962
Median	3.315
Mode	3
Standard Deviation	0.386812669
Standard Error	0.054703572
Sample Variance	0.149624041
Kurtosis	3.141968267
Skewness	-1.456820799
Range	1.82
Minimum	2
Maximum	3.82
Sum	164.81
Count	50
Confidence Level (95%)	0.107217031
z-test	-0.069465299
P-value	0.472309627

Step 3:

P-value = 0.4723

Step 4:

Since (P-value = 0.4723) > (α = 0.05), do not reject H_0.

Step 5:

At α = 0.05, there is enough evidence to support the claim that the CGPA of UMPSA students that study less than two hours a day is at least 3.30.

7.0 HYPOTHESIS TESTING FOR TWO POPULATION MEANS

b. Choose one probability sampling method to select less than 30 data from each group.

i. Identify which sampling method you choose to select the data and explain the sampling method process. (2 Marks)

Random sampling method is chosen to select the data. For this sampling method, there is an equal chance for each data to be chosen. First, all the data collected from each group are listed and labelled from 1 until 50. Then, using Microsoft Excel, a total of 15 samples were selected randomly for each group.

ii. Present the selected data in a table. (2 Marks)

Table 11: Stratified Random Sampling of Group 1 (less than 2 hours) and Group 2 (more than 2 hours)

Group 1		Group 2	
Less Than Two Hours	CGPA	More Than Two Hours	CGPA
35	3.5	6	3.1
29	3.46	23	3.3
45	3.7	32	3
37	3.57	1	2.45
18	3.2	12	4
31	3.5	42	3.7
10	3	8	3.8
9	3	40	3.6
26	3.33	10	3.9
34	3.5	4	3.5
21	3.3	33	3.6
19	3.22	9	3.66
16	3.2	28	3.8
22	3.3	31	3.5
3	2.49	14	3.94

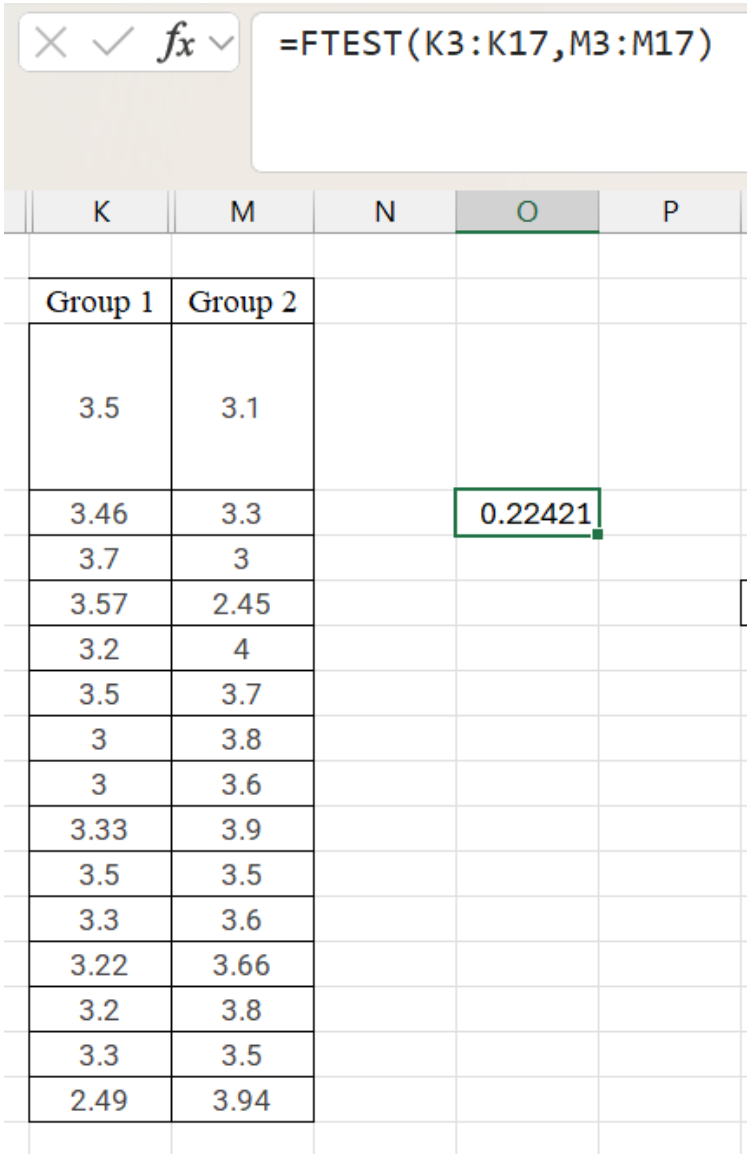
c. Create a situation to conduct a hypothesis testing using the data selected in (b) to compare two population means between the groups. (12 Marks)

Since the total population mean for the CGPA of UMPSA students is not available, it is assumed that students who study more than 2 hours a day have a higher CGPA than those who study less than 2 hours.

To conduct a hypothesis testing to compare two population means between two groups, the population variances, σ_1^2 and σ_2^2, must first be assessed to determine whether they are equal or not. Since the groups are based on different study durations (less than 2 hours and 2 hours or more), it is assumed that the population variances of CGPA for these two groups are not equal.

Step 1: (a) $H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$ (claim)

(b) Obtain P-value using F-Test Command in Microsoft Excel.



P-value = 0.2242062.

(c) Since (P-value = 0.2242062) > (α = 0.05), then we accept H_0.

(d) At = 0.05, there is enough evidence to reject the claim that the population variances of CGPA for these two groups are not equal. Therefor, there is no difference in variability.

Step 2: Claim: Students who study more than 2 hours (Group 2) has higher CGPA than students who study less than 2 hours (Group 1).

$H_0: \mu_2 - \mu_1 \leq 0$
$H_1: \mu_2 - \mu_1 > 0$ (claim)

Step 3: T-test and Microsoft Excel Output.

t-Test: Two-Sample Assuming Equal Variances

	Group 1	*Group 2*
Mean	3.284666667	3.523333333
Variance	0.088455238	0.172395238
Observations	15	15
Pooled Variance	0.130425238	
Hypothesized Mean Difference	0	
df	28	
t Stat	-1.809845825	
P(T<=t) one-tail	0.040533268	
t Critical one-tail	1.701130934	
P(T<=t) two-tail	0.081066535	
t Critical two-tail	2.048407142	

Step 4: Since the test is left-tailed test, hence P-value = 0.0405

Step 5: Since (P-value = 0.0405) < (α = 0.05), then we reject H_0.

Step 6: At $\boldsymbol{\alpha}$ = 0.05, there is enough evidence to support the claim that students who study more than 2 hours has higher CGPA than students who study less than 2 hours.

8.0 CONCLUSION

Based on your problem/topic stated in (1), give any relevant conclusion for the study. (1 Mark)

In conclusion, the CGPA of the UMPSA students that study more than 2 hours is more than the students that study less than 2 hours. Thus, this trend strongly suggests that increasing study time results in higher CGPA, reflecting better academic performance.

9.0 REFERENCES

Bradburn, S. (2021). *How To Box Plot in* Retrieved from Youtube: https://www.youtube.com/watch?v=39lsUsJsc2c

Lanoue, S. (2025). *How to Add P Value to Graph in Excel*. Retrieved from Bricks: https://www.thebricks.com/resources/guide-how-to-add-p-value-to-graph-in-excel

Maryam, L., Tuba, A., & Dr. Ashan, N. (2025). The Impact of Screen Time on Students Learning in Higher Education. *Dialogue Social Science Review (DSSR)*, 330–338.

Tan, T. H., & Yap, J. K. (2024). The Impact of CGPA Goal Setting on Student Performance. *The European Conference on Education 2024*, (pp. 1-8).

Descriptive statistics in Excel. (n.d.). Step by Step Tutorial. https://www.excel-easy.com/examples/descriptive-statistics.html

APPENDICES