اونيۏرسيتي مليسيا ڤهڠ UNIVERSITI MALAYSIA PAHANG PUSAT SAINS MATEMATIK	**SEMESTER II 2021/2022** **BUM 2413 APPLIED STATISTICS** **ASSIGNMENT** **TOTAL MARK: 60**	**DUE DATE: 03/06/2022**

The objectives of this assignment are to help you to understand the statistical analysis process, to analyse data using software and to develop your integrity in reporting your assignment. The best way to understand statistics is by involving yourself in the whole statistical process and not just limited to studying statistics from books, videos, or websites. This assignment requires you to follow the steps of statistical problem-solving methodology by conducting your own study. You will experience on how to collect, organise, summarise, analyse, present, interpret, and draw conclusion from data, as well as preparing a report of your study.

INSTRUCTIONS:

1. Set up a group that consists of **four (4)** or **five (5)** members from your section only and name your group using any **statistical term.**

2. Obtain an **APPROVAL** of your chosen topic from your lecturer **BEFORE** you start collecting data and begin your statistical analysis.

3. Use the template on **page 5** as the cover for your assignment booklet. Fill in all the required particulars clearly.

4. Answer **ALL** questions and use **appropriate statistical notations**.

5. Perform **ALL** analysis using *Microsoft Excel* and *P*-value approach.

6. **Submit** the following items for **EACH group:**
   (i) **A softcopy report (pdf file)** that includes all attachments of any relevant evidences (*Microsoft Excel* outputs, handwritten data record, photos, Google Docs, and etc.) in the appendix section.
   (ii) **A softcopy** of the Excel worksheet that you worked on to complete the report.
   (iii)**Submit via KALAM**, name your file as (Section)_(group name), for example: 01G_MEAN.

7. **LATE submission of assignment will not be entertained.**

1. **Identify** a problem that you are interested to study. Provide a brief **description** of your study. (2 Marks)

   What is the time spent on exercise among IPTA university student. The survey is made for collecting the data about time spent on exercise in a week among IPTA students. We want to investigate the average hours of exercise per week spent by the university student. The respondents are required to answer a few questions related to our study.

2. Choose a **single quantitative variable** that describe your chosen problem. Identify the type of level of measurement for the variable. (2 Marks)

   single quantitative variable : Time Spent on Exercise per week (hours).

   level of measurement : Ratio level

3. State your **population**. (1 Mark)

   IPTA university student

4. Divide the data collected into **two significant groups** (e.g.: gender, faculty, year of study, etc.) that related to the study. Please make sure you have **at least 50 respondents** for each group.

   (i) State the **name of the groups**.

   Males(Group 1) and Female (Group 2)

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

Group 1	**Group 2**
0	*0*
0.5	*0*
0.75	*0.25*
1	*0.5*
1	*0.5*
1	*1*
1	*1*
1.5	*1*
1.5	*1*
2	*1*
2	*1*
2	*1*
2	*1*
2	*1*
2	*1*
3	*1*
3	*2*
3	*2*
3	*2*
3	*2*
3	*2*
3	*2*
3	*2*
3	*2*
3	*2*
4	*3*
4	*3*
4	*3*
4	*3*
4	*4*
4	*4*
5	*4*
5	*4*
6	*4*
6	*4*
6	*4*
7	*4*
7	*4*
7	*5*
7	*5*
8	*5*
8	*5*
8	*5*
8	*5*
10	*5*
10	*6*

10 *10* *12* *14* *15* *30*	*6* *7* *7* *7* *7* *8* *8* *8* *9*

(2 Marks)

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

Questionnaires and survey. Evidence :

(iv) Identify which **sampling method** you used to collect the data. **Explain** the sampling method process.

Convenien

5. For each set of data, obtain the **descriptive statistics** using *Microsoft Excel*. Then, summarise the **measures of central tendency** and **measures of variation** in the following table. (3 Marks)

Group Name	Measures of central tendency	Measures of variation
Group 1	Mean = 5.2548 Median = 4 Mode = 3 Midrange = 15	Range = 30 Variance = 24.6752 Standard deviation = 4.9674 Coefficient of Variation = 94.5307%
Group 2	Mean = 3.3864 Median = 3 Mode = 1 Midrange = 4.5	Range = 9 Variance = 5.7936 Standard deviation = 2.4070 Coefficient of Variation = 71.0784%

6. **Compare** and **comment** the measures of central tendency and measures of variation between **Group 1** and **Group 2**.

(4 Marks)

In the measures of central tendency, the mean of Group 1(Male) data is larger than Group 2(Female), thus in average Group 1(Male) spent more time on exercise per week.

the distribution of data for Group 1(Male) is right-skewed or positively skewed since its Mean > Median > Mode from the table above, which the data is almost cluster at the left side of the distribution shape. Meanwhile the distribution of data for Group 2(Female) is also right skewed or positively skewed since its Mean > Median > Mode from the table above, which the data is almost cluster at the left side of the distribution shape.

In the measures of variation, since the standard deviations of Group 1(Male) is larger than Group 2(Female), therefore Group 1(Male)'s data is more dispersed, less consistent, and less precise as compared to Group 2(Female).

The skewness value of Group 1(Male) is higher than the Group 2(Female). Therefore the distribution of data from Group 1(Male) is more skewed to the left, means that Group 1(Male) spent more time on exercise per week.

7. Construct **boxplots** for the four sets of data on the **same axis**. Identify the **shape of distribution** for each boxplot. **Compare** and **comment** on the average and variability of the boxplots.

(8 Marks)

Based on the location of median, Group 1(Male) has right-skewed distribution where most the time spent is concentrated at the time of less than 4 hours since the median falls to the below of the centre of the box. However, Group 2(Female) has also right-skewed distribution where most the time spent is concentrated at the time of less than 3 hours since the above line is longer than the below line when the box is approximately symmetric.

For average, based on the median value, 50% of time spent by male is less than 4 hours, whereas 50% of time spent by female is less than 3 hours.

For IQR of Group 1(Male) is 5 hours where most 50% of time spent by male is between 2 to 7 hours. Meanwhile, for IQR of Group 2(Female) is 4 hours where most 50% of time spent by female is between 1 to 5 hours. Hence, the variation of time spent on exercise by male is higher than time spent on exercise by female since (IQR of male = 5 hours) > (IQR of female = 4 hours).

For data of Group 1(male), since the lower boundary and upper boundary are -5.5 and 14.5 respectively, thus there is having two outliers which are 15 hours and 30 hours. For data of Group 2(female), since the upper boundary and upper boundary are -5 and 11 respectively, thus there is no outlier exists in this data set.

8. What is the **best measure of central tendency** to describe your data? Give a **reason**.

(2 Marks)

The best measure of central tendency in this data is using Median as the distribution is right-skewed or positively skewed.

9. What is the **best measure of variation** to describe your data? Give a **reason**.

(2 Marks)

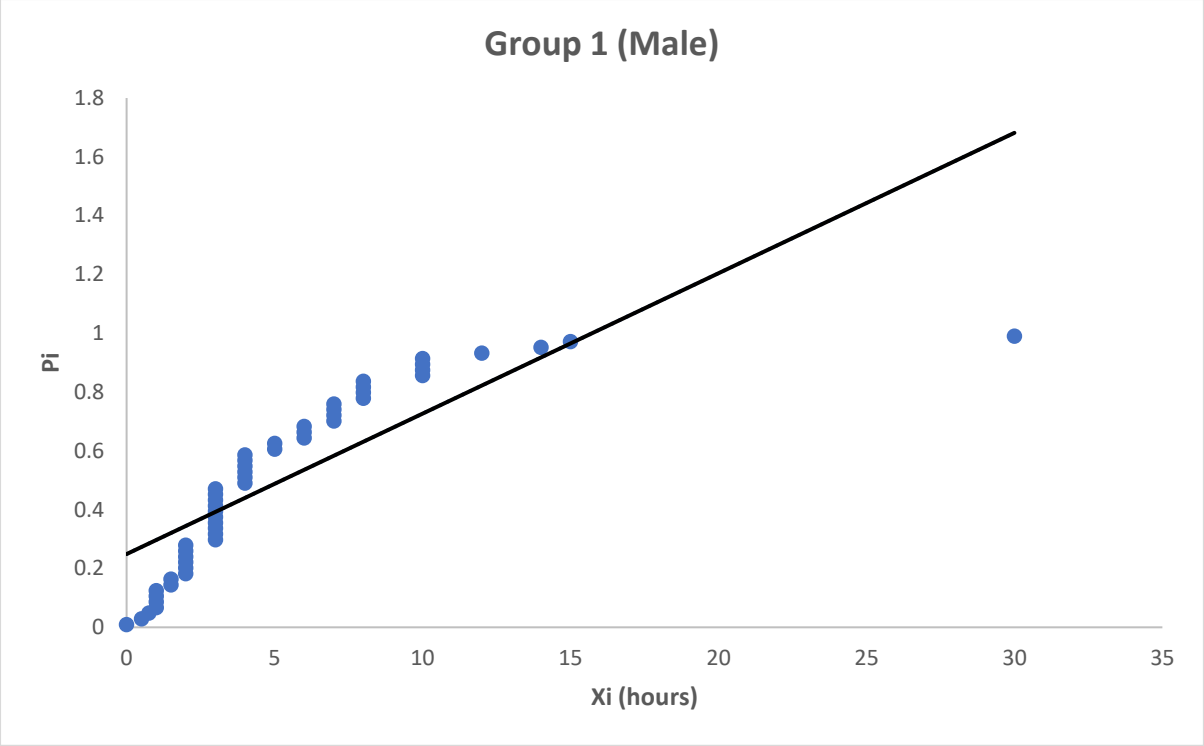
The best measure of variation in this data is using Interquartile range (IQR) as the distribution data of these both of groups is right-skewed or positively skewed.

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

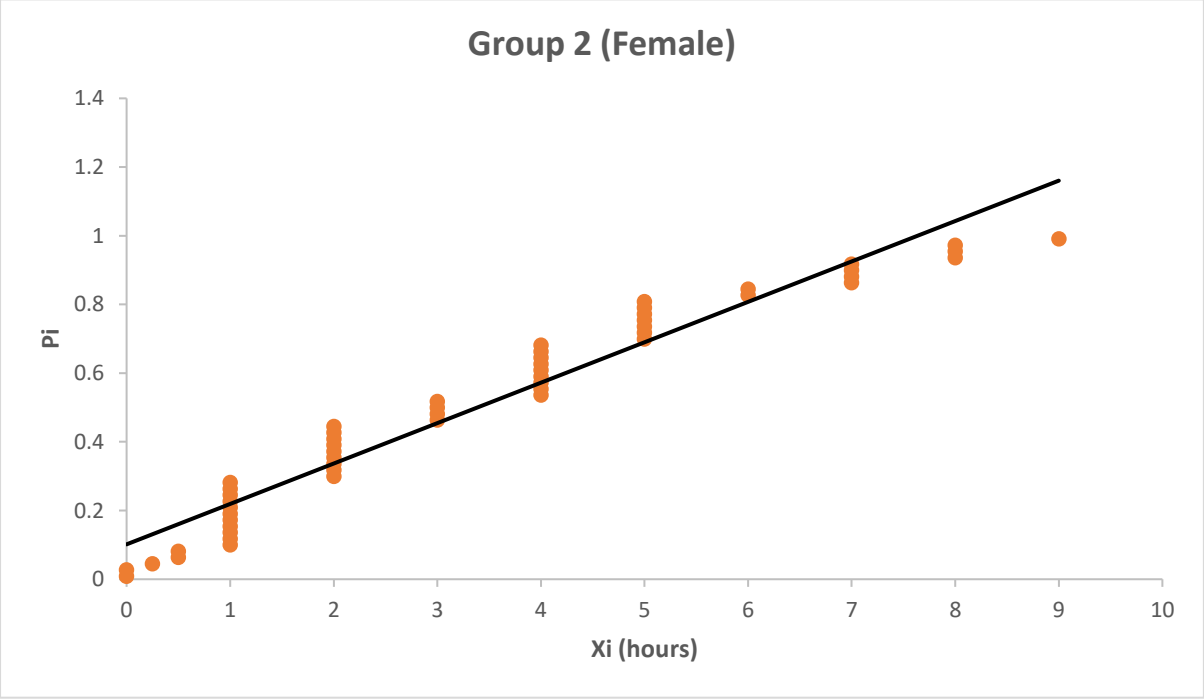
(4 Marks)

Group 1			Group 2		
i	Time Spent on Exercise per week	**pi = (i - 0.5) / n**	***i***	Time Spent on Exercise per week	**pi = (i - 0.5) / n**
1	0	0.009615385	1	0	0.009090909
2	0.5	0.028846154	2	0	0.027272727
3	0.75	0.048076923	3	0.25	0.045454545
4	1	0.067307692	4	0.5	0.063636364
5	1	0.086538462	5	0.5	0.081818182
6	1	0.105769231	6	1	0.1
7	1	0.125	7	1	0.118181818
8	1.5	0.144230769	8	1	0.136363636
9	1.5	0.163461538	9	1	0.154545455
10	2	0.182692308	10	1	0.172727273
11	2	0.201923077	11	1	0.190909091
12	2	0.221153846	12	1	0.209090909
13	2	0.240384615	13	1	0.227272727
14	2	0.259615385	14	1	0.245454545
15	2	0.278846154	15	1	0.263636364
16	3	0.298076923	16	1	0.281818182
17	3	0.317307692	17	2	0.3
18	3	0.336538462	18	2	0.318181818
19	3	0.355769231	19	2	0.336363636
20	3	0.375	20	2	0.354545455
21	3	0.394230769	21	2	0.372727273
22	3	0.413461538	22	2	0.390909091
23	3	0.432692308	23	2	0.409090909

24	3	0.451923077	24	2	0.427272727
25	3	0.471153846	25	2	0.445454545
26	4	0.490384615	26	3	0.463636364
27	4	0.509615385	27	3	0.481818182
28	4	0.528846154	28	3	0.5
29	4	0.548076923	29	3	0.518181818
30	4	0.567307692	30	4	0.536363636
31	4	0.586538462	31	4	0.554545455
32	5	0.605769231	32	4	0.572727273
33	5	0.625	33	4	0.590909091
34	6	0.644230769	34	4	0.609090909
35	6	0.663461538	35	4	0.627272727
36	6	0.682692308	36	4	0.645454545
37	7	0.701923077	37	4	0.663636364
38	7	0.721153846	38	4	0.681818182
39	7	0.740384615	39	5	0.7
40	7	0.759615385	40	5	0.718181818
41	8	0.778846154	41	5	0.736363636
42	8	0.798076923	42	5	0.754545455
43	8	0.817307692	43	5	0.772727273
44	8	0.836538462	44	5	0.790909091
45	10	0.855769231	45	5	0.809090909
46	10	0.875	46	6	0.827272727
47	10	0.894230769	47	6	0.845454545
48	10	0.913461538	48	7	0.863636364
49	12	0.932692308	49	7	0.881818182
50	14	0.951923077	50	7	0.9
51	15	0.971153846	51	7	0.918181818
52	30	0.990384615	52	8	0.936363636
			53	8	0.954545455
			54	8	0.972727273
			55	9	0.990909091



The data appear of group 1 does not come from an approximately normal distribution. Since the data does not lies approximately on a straight line, the data are not normally distributed.



11. In Chapter 3, you have learned statistical hypothesis concerning a parameter(s) of one and two populations. The parameters are the population mean, proportion, variance and standard deviation. Hypothesis testing is one of the inferential statistics in statistical analysis. Assuming that the data obtained in **(5)** is normally distributed population, answer the following questions using **P-value approach and *Microsoft Excel***.

    (**NOTE**: Create your own hypothesised mean with justification)

    a. Conduct a **hypothesis testing for one population mean** from one of the groups.

(8 Marks)

A sport company is studying the weekly exercise behaviour of female student in order to balance the nutrient of their female nutrition supplement product. The company estimate the time spent on exercise weekly is 3 hours . 55 female IPTA students submitted their weekly exercise record as shown below. Do you think the sport company over estimate the time spent on exercise weekly of female ?

FEMALE
0
0
0.25
0.5
0.5
1
1
1
1
1
1
1
1
1
1
1
2
2
2
2
2
2
2
2
2
2
3
3
3

3
4
4
4
4
4
4
4
4
4
5
5
5
5
5
5
5
6
6
7
7
7
7
8
8
8
9

Step 1: Hypothesis
H_0: $\mu \geq 3$
H_1: $\mu < 3$ (claim: over estimate the time spent of exercise weekly)

Step 2:

FEMALE	
Mean	3.386364
Standard Error	0.324557
Median	3
Mode	1
Standard Deviation	2.406982
Sample Variance	5.793561
Kurtosis	-0.66773
Skewness	0.547314
Range	9
Minimum	0
Maximum	9
Sum	186.25
Count	55
Confidence Level(95.0%)	0.650698

z-test	1.190435
P-value	0.883062

Step 3: Calculate t-test/z-test
$n \geq 30$, so use z-test.
P-value= 0.8831

Step 4:
Since (P-value=0.8831)>(α=0.05), then we do not rejectH_0.

Step 5:
At α=0.05, there is insufficient evidence to support the claim that the sprt company over estimate the time spent of exercise weekly of female.

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

i. Identify which sampling method you used to select the data. Explain the sampling method process.

(2 Marks)

Stratified sampling.

ii. Present the selected data in a table.

Group 1	**Group 2**
2	*2*
2	*3*
14	*2*
6	*5*
7	*2*
1.5	*7*
10	*2*
8	*4*
3	*5*
2	*4*
4	*1*
10	*0.5*
3	*4*
12	*4*
7	*4*

(2 Mark)

c. Use the data selected in (b) to conduct a **hypothesis testing to compare two population means** between the two groups.

(12 Marks)

A company is considering developing a new outfit product depending on time spent on exercise weekly record by male and female. The company will use the more fast fry material for male outfit if the time spent of exercise weekly of male is greater than female. Will the more fast dry material be used in the male outfit?

Step 1:
(i) H_0: $\sigma_M^2 = \sigma_F^2$
H_1: $\sigma_M^2 \neq \sigma_F^2$ (claim: significant difference in variability)

(ii)

P-value 0.003117922

(iii) Since (P-value = 0.0031) < (α = 0.05), we reject H0.

(iv) At $\alpha = 0.05$, there is enough evidence to support the claim. Therefore, there is difference in variability. $\sigma_M^2 \neq \sigma_F^2$

Step 2:
H_0: $\mu_M - \mu_F \leq 0$
H_1: $\mu_M - \mu_F > 0$ (claim: More fast dry material will be used)

Step 3:
Choose the t-Test: Two-Sample Assuming Equal Variances

t-Test: Two-Sample Assuming Unequal Variances

	Male	*Female*
Mean	6.1	3.3
Variance	16.29285714	2.992857143
Observations	15	15
Hypothesized Mean Difference	0	
df	19	
t Stat	2.46936789	
P(T<=t) one-tail	0.011593821	
t Critical one-tail	1.729132812	
P(T<=t) two-tail	0.023187642	
t Critical two-tail	2.093024054	

Step 4:
The test is right-tailed test, hence P-value = 0.0116

Step 5:
Since (P-value = 0.0116) < ($\alpha = 0.05$), then we rejectH_0.

Step6:
At $\alpha = 0.05$, there is enough evidence to support the claim. Therefore, the more dry fast material will be used in male outfit.

12. Based on your problem stated in (1), give any relevant **conclusion** for the study

(2 Marks)

The male time spent is greater than the female.