UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY & STATISTICAL DATA ANALYSIS

SECI2143

Lecturer:

Rozilawati Dollah @ Md Zain

Group Members:

No	Name	Matric No
1	MUHAMAD AMSYAR BIN IBRAHIM	A21EC0058
2	AIMAN HAIKAL BIN ZAINUDDIN	A21EC0154
3	MUHAMMAD DANIAL WAJDI BIN SAFIAY	A21EC0071
4	ARSYAD WIRATAMA	A20EC0290

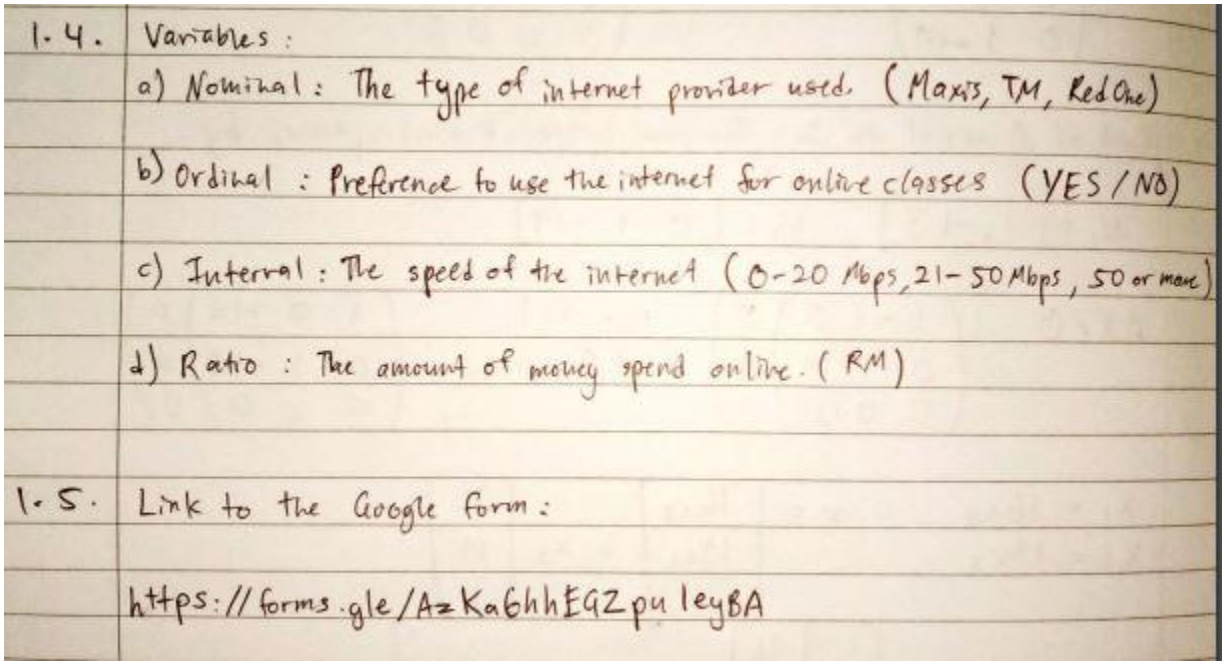
Question 1:

1.1 Related to the study, descriptive statistics is used to summarize the characteristics of the data set population of the students in Malaysia based on the data received from the response of the computer science students. On the other hand, inferential statistics enable us to test hypothesis regarding the population based on the sample of Computer Science student in order to predict some events in the future concerning the population

1.2. As the respondent is Computer Science students in Malaysia on a survey on the usage of internet facilities, so we can assume that Computer Science students are the sample of the study while the whole student body of Malaysia or even rather the whole population of the country can be accept as the population of the study.

1.3. To analyse the data regarding the study, first we need to produce a survey with questions on the usage of internet facilities by the Computer Science students. ~~After the~~ Then, the survey must be spread in order to received appropriate amount of respondents for the data to be analysed correctly. After the response has been received, we need to tabulate the data into tables and chart in order to see the pattern regarding the study. Through the result of the survey on the Computer Science students, we can build a hypothesis on how the population behavior on the usage of internet facilities relate to the response by the Computer Science students.

During the collection of data, there is a data cleaning process where we remove duplicate records, errors on the survey and white spaces in order to provide an accurate findings on the sample.



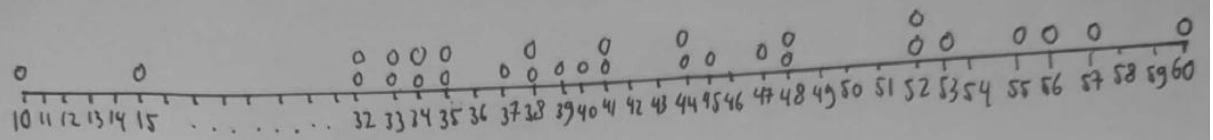
Link to the Google form : https://forms.gle/AzKa6hhEGZpu1eyBA

Question 1

10, 15, 32, 32, 33, 33, 34, 34, 35, 35, 37, 38, 38, 39, 40, 41, 41, 44, 44, 45, 47, 48, 48, 52, 52, 53, 55, 56, 57, 60

2.1) Represent the data form of:

a) Dot Plot



b.) Stem and Leaf Plot

Stem	Leaf
1	0 5
2	
3	2 2 3 3 4 4 5 5 7 8 8 9
4	0 1 1 4 4 5 7 8 8
5	2 2 3 5 6 7
6	0

2.2) If you are the second fastest, what percentile are you at?

$Y(K) = 15$

percentile of value 15 $= \frac{\text{number of value less than 15}}{\text{total number of values}} \times (100) = \frac{1}{30}(100) = 3.33$

The second fastest student (15) is at the 3.33 percentile

Question 3

3.1)

Class interval	Class boundaries	Midpoint	Frequency, f	Cumulative frequency
0.5 - 9.5	0 - 10	5	7	7
10.5 - 19.5	10 - 20	15	10	17
20.5 - 29.5	20 - 30	25	23	40
30.5 - 39.5	30 - 40	35	51	91
40.5 - 49.5	40 - 50	45	6	97
50.5 - 59.5	50 - 60	55	3	100

3.3) mean, $\bar{x} = \frac{5(7) + 15(10) + 25(23) + 35(51) + 45(6) + 55(3)}{100}$

$= \frac{2980}{100}$

$= 29.8$

median $= L + \left(\frac{N/2 - cf_p}{f_{med}}\right) w$

$= 30.5 + \left(\frac{50 - 40}{51}\right)(10)$

$= 32.4608$

mode $= l + h \times \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)}$

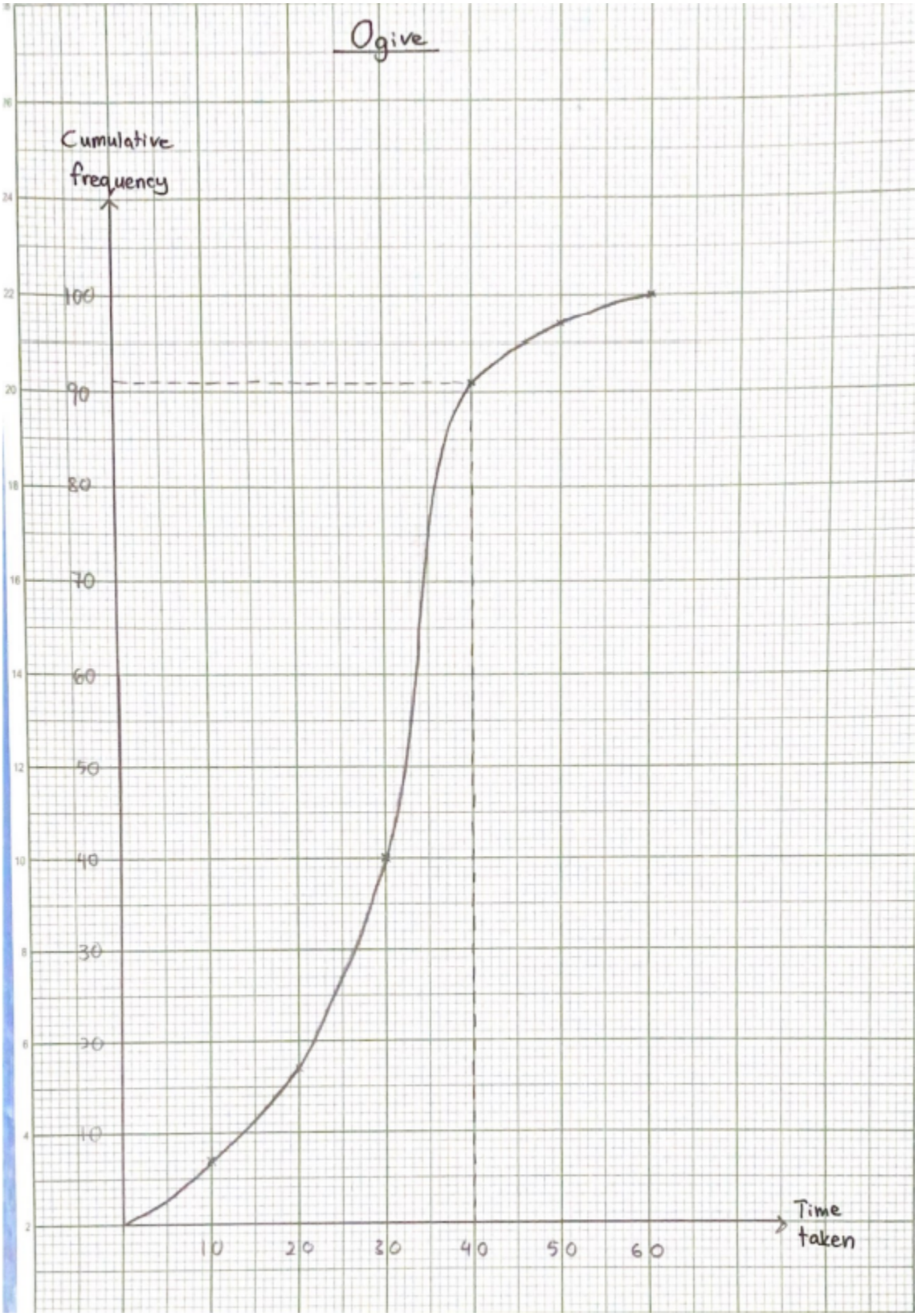
$= 30.5 + 10 \times \frac{(51 - 23)}{[2(51) - 23 - 6]}$

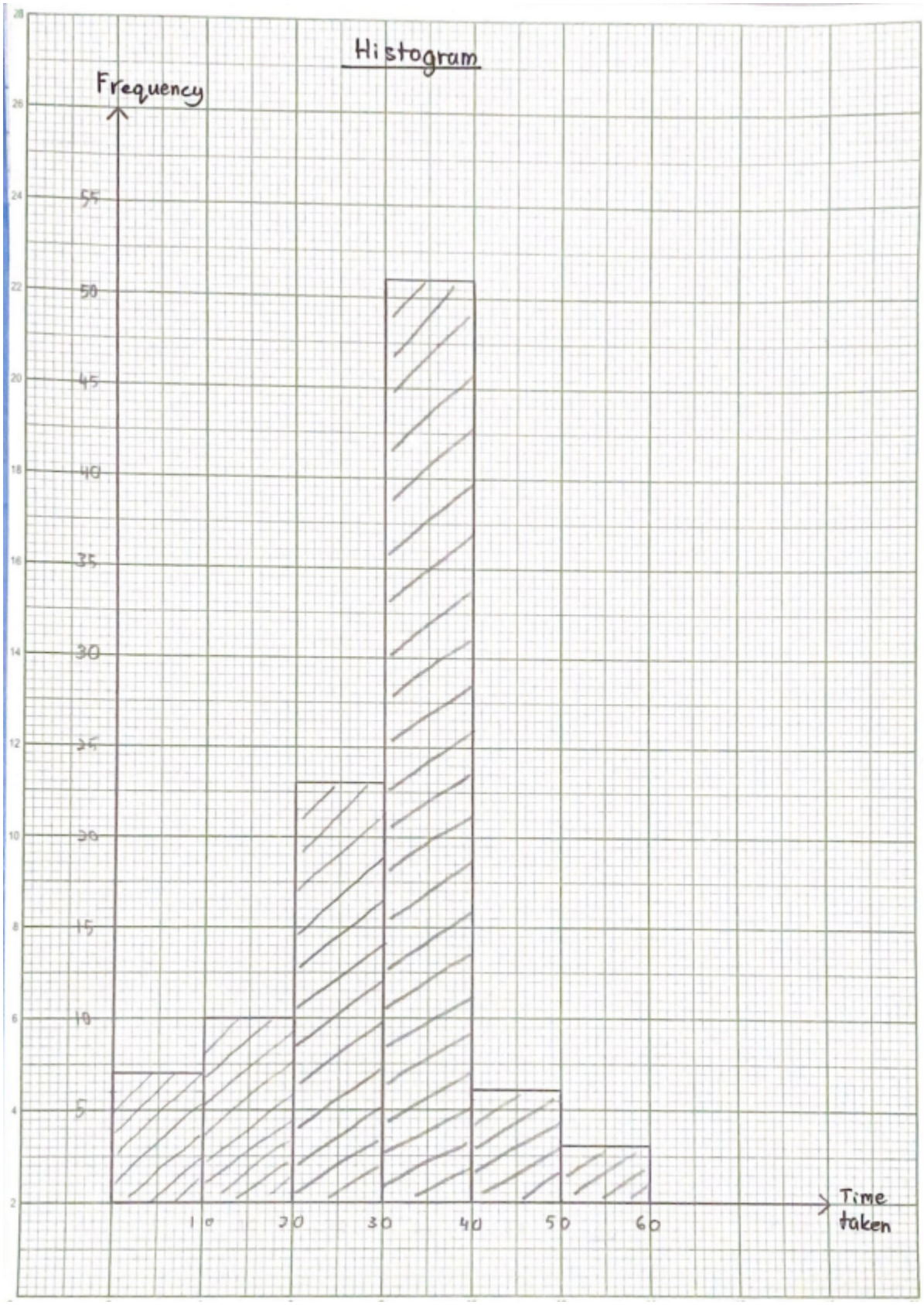
$= 30.5 + 10 \times \frac{28}{73}$

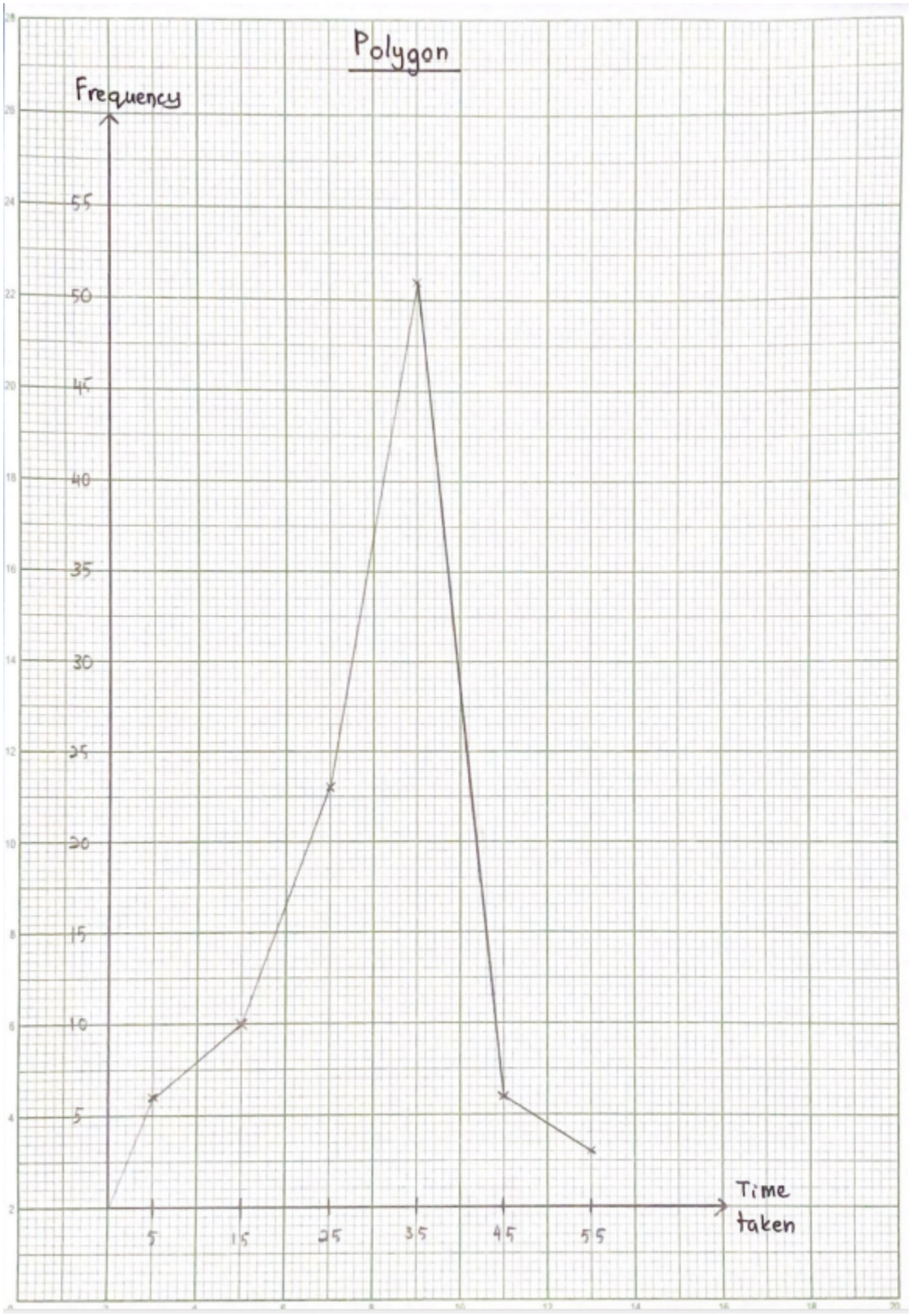
$= 34.3356$

3.2) d. percentile of students answer the assessment >40 minutes $= \frac{91}{100} \times 100$

$= 91^{th}$ percentile


Ogive
Cumulative frequency
100
90
80
70
60
50
40
30
20
10
10
20
30
40
50
60
Time taken


Histogram
Frequency
55
50
45
40
35
30
25
20
15
10
5
10
20
30
40
50
60
Time taken


Polygon
Frequency
55
50
45
40
35
30
25
20
15
10
5
Time
taken
5
15
25
35
45
55

Assignment 1

Question 4

4.1) For both datasets, produce summary statistics which compromise sample

UNIVERSITI CANGGIH MALAYSIA sample of data

a) range

Max value - Min Value = 60 - 10 = 50

b) mean

$\bar{x} = \frac{\Sigma x}{N} = \frac{1228}{30} = 40.93$

c) median

Data: 10, 15, 32, 32, 33, 33, 34, 34, 35, 35, 37, 38, 38, 39, 40, 41, 41, 44, 44, 45, 47, 48, 48, 52, 52, 53, 55, 56, 57, 60

(10, 15, 20, 23)

$\frac{30}{2} = 15 \rightarrow 15^{th}$ & 16^{th} Data

$\text{Median} = \frac{40 + 41}{2} = 40.5$

d) mode

mode = 32, 33, 34, 35, 38, 41, 44, 48, 52

e) first Quartile, $Q_1 = P_{25}$

$i = \frac{P_{25}}{100}(N)$

$= \frac{25}{100}(30)$

$= 7.5 = 8$

$Y[8] = 34$

$Q_1 = 34$

f) third Quartile, $Q_3 = P_{75}$

$i = \frac{P_{75}}{100}(N)$

$= \frac{75}{100}(30$

$= 22.5$

$= 23$

$Y[23] = 48$

$Q_3 = 48$

g) Interquartile range

Interquartile range $= Q_3 - Q_1$

$= 48 - 34 = 14$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
10	-30.93	956.6649	-29589.64	915207.73
15	-25.93	672.3649	-17434.42	452074.55
32	-8.93	79.7449	-712.12	6359.24
32	-8.93	79.7449	-712.12	6359.24
33	-7.93	62.8849	-498.67	3954.51
33	-7.93	62.8849	-498.67	3954.51
34	-6.93	48.0249	-332.81	2306.39
34	-6.93	48.0249	-332.81	2306.39
35	-5.93	35.1649	-208.52	1236.57
35	-5.93	35.1649	-208.52	1236.57
37	-3.93	15.4449	-60.69	238.54
38	-2.93	8.5849	-25.15	73.7
38	-2.93	8.5849	-25.15	73.7
39	-1.93	3.7249	-7.18	13.87
40	-0.93	0.8649	-0.8	0,74
41	0.07	0.0049	$343 \cdot 10^{-6}$	0.000024
41	0.07	0.0049	$343 \cdot 10^{-6}$	0.000024
44	3.07	9.4249	28.93	88.82
44	3.07	9.4249	28.93	88.82
45	4.07	16.5649	67.41	274.39
47	6.07	36.8449	223.64	1357.54
48	7.07	49.9849	353.39	2498.49
48	7.07	49.9849	353.39	2498.49
52	11.07	122.5449	1356.57	15017.25
52	11.07	122.5449	1356.57	15017.25
53	12.07	145.6849	1758.41	21224.09
55	14.07	197.9649	2785.36	39190.1
56	15.07	227.1049	3422.47	51576.63
57	16.07	258.2449	4149.99	66690.42
60	19.07	363.6649	6935.08	132252.15

Sum of

$(x_i - \bar{x})^2$

↳ 3727.867

$(x_i - \bar{x})^3$

↳ -27618.6

$(x_i - \bar{x})^4$

↳ 1743170.66

h) Variance

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$= \frac{3727.86}{30-1}$$

$$= 128.54$$

i) Standard Deviation

$$S = \sqrt{S^2}$$

$$= \sqrt{128.54}$$

$$= 11.33$$

j) Skewness

$$\text{Skewness} = \frac{\sum (x_i - \bar{x})^3}{(N-1)S^3}$$

$$= \frac{-27618.6}{(30-1)(11.33)^3}$$

$$= \frac{-27618.6}{(29)(1454.41)}$$

$$= \frac{-27618.6}{42177.89} = -0.65$$

Skewness < 0

∴ Therefore, the distribution is asymmetrical and points in the negative direction

k) Kurtosis

$$\text{Kurtosis} = \frac{\sum (x_i - \bar{x})^4}{(N-1)S^4}$$

$$= \frac{1743170.66}{(30-1)(11.33)^4}$$

$$= \frac{1743170.66}{47787.53} = 3.64$$

→ the kurtosis > 0, therefore it is Leptokurtic

No.: Date:

For UNIVERSITI MUHIBBAH MALAYSIA

Data: 48, 48, 49, 49, 49, 50, 50, 50, 51, 52, 52, 53, 54, 54, 54, 55, 55, 56, 57, 57, 57, 59, 59, 60, 63, 65, 70, 72, 73, 90

4.1) a- Range

Range = largest value - smallest

= 90 - 48

= 42

b- Mean

$$\bar{x} = \frac{48+48+49+49+49+50+50+50+51+52+52+53+54+54+54+55+55+56+57+57+57+59+59+60+63+65+70+72+73+90}{30}$$

$$= \frac{1711}{30}$$

= 57.03

c- Median

$$\text{Median} = \frac{54+55}{2} = 54.5$$

d- Mode

Mode = 49, 50, 54, 57

e- first Quartile, $Q_1 = P_{25}$

$$i = \frac{P_{25}(N)}{100}$$

$$= \frac{25\ (30)}{100}$$

= 7.5

= 8

∴ Y[8] = 50

Q_1 = 50

No.: Date:

f - third Quartile, $Q_3 = P_{75}$

$i = \frac{P_{75}}{100}(N)$

$= \frac{75}{100}(30)$

$= 22.5$

$= 23$

$\therefore Y[23] = 59$

$Q_3 = 59$

g - Interquartile range

Interquartile range $= Q_3 - Q_1$

$= 59 - 50$

$= 9$

No.: Date:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
48	-9.03	81.5409	-736.31	6648.92
48	-9.03	81.5409	-736.31	6648.92
49	-8.03	64.4809	-517.78	4157.79
49	-8.03	64.4809	-517.78	4157.79
49	-8.03	64.4809	-517.78	4157.79
50	-7.03	49.4209	-347.43	2442.43
50	-7.03	49.4209	-347.43	2442.43
50	-7.03	49.4209	-347.43	2442.43
51	-6.03	36.3609	-219.26	1322.12
52	-5.03	25.3009	-127.26	640.14
52	-5.03	25.3009	-127.26	640.14
53	-4.03	16.2409	-65.45	263.77
54	-3.03	9.1809	-27.82	84.29
54	-3.03	9.1809	-27.82	84.29
54	-3.03	9.1809	-27.82	84.29
55	-2.03	4.1209	-8.37	16.98
55	-2.03	4.1209	-8.37	16.98
56	-1.03	1.0609	-1.09	1.13
57	-0.03	0.0009	-2.7×10^{-5}	8.1×10^{-7}
57	-0.03	0.0009	-2.7×10^{-5}	8.1×10^{-7}
57	-0.03	0.0009	-2.7×10^{-5}	8.1×10^{-7}
59	1.97	3.8809	7.65	15.06
59	1.97	3.8809	7.65	15.06
60	2.97	8.8209	26.20	77.81
63	5.97	35.6409	212.78	1270.27
65	7.97	63.5209	506.26	4034.90
70	12.97	168.2209	2181.83	28298.27
72	14.97	224.1009	3354.79	50221.21
73	15.97	255.0409	4073.00	65045.86
90	32.97	1087.0209	35839.08	1181614.44
		=2494.9670	=41500.45	1366845.47

$\Sigma(x_i - \bar{x})^2 = 2494.9670$, $\Sigma(x_i - \bar{x})^3 = 41500.45$, $\Sigma(x_i - \bar{x})^4 = 1366845.47$

No.: Date:

h - variance

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

$$= \frac{2494.9670}{30-1}$$

$$= 86.03$$

i - Standard deviation

$$s = \sqrt{s^2}$$

$$= \sqrt{86.03}$$

$$= 9.28$$

j - Skewness

$$\text{Skewness} = \frac{\Sigma (x_i - \bar{x})^3}{(N-1)s^3}$$

$$= \frac{41500.45}{(30-1)(9.28)^3}$$

$$= 1.79$$

∴ Skewness > 0

Therefore, the distribution is asymmetrical and points in the positive direction.

k - Kurtosis's

$$\text{kurtosis} = \frac{\Sigma (x_i - \bar{x})^4}{(N-1)s^4}$$

$$= \frac{1366845.47}{(30-1)(9.28)^4}$$

$$= 6.36$$

∴ since kurtosis > 0, it is Leptokurtic

No.: Date:

4.2) For UNIVERSITI CANGGIH MALAYSIA

$= Q_3 + 1.5(\text{Interquartile range})$ $= Q_1 - 1.5(\text{Interquartile range})$

$= 48 + 1.5(14)$ $= 34 - 1.5(14)$

$= 69$ $= 13$

∴ There is outlier for UNIVERSITI CANGGIH MALAYSIA since there is data value that is not between 13 to 69.

For UNIVERSITI MUHIBBAH MALAYSIA

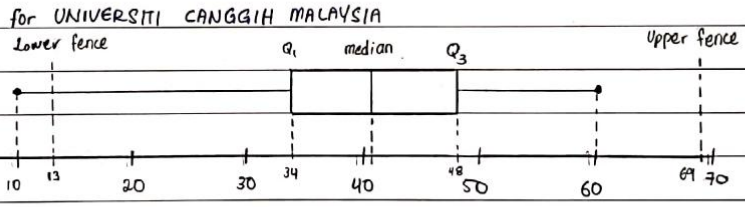
$= Q_3 + 1.5(\text{Interquartile range})$ $= Q_1 - 1.5(\text{Interquartile range})$

$= 59 + 1.5(9)$ $= 50 - 1.5(9)$

$= 72.5$ $= 36.5$

∴ There are outlier for UNIVERSITI MUHIBBAH MALAYSIA since there exist data value that are not between 36.5 to 72.9.

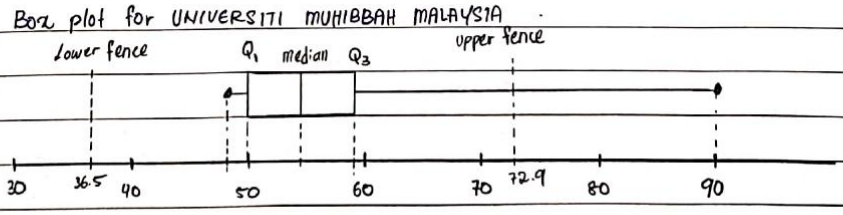
4.3) Box plot for UNIVERSITI CANGGIH MALAYSIA



maximum value = 60 Upper fence = 69 Q_3 = 48 median = 40.5

minimum value = 10 Lower fence = 13 Q_1 = 34

Box plot for UNIVERSITI MUHIBBAH MALAYSIA



maximum value = 90 Upper fence = 72.9 Q_3 = 59 median = 54.5

minimum value = 48 Lower fence = 36.5 Q_1 = 50

4.4) Comment on the differences between the two datasets in terms of

a) Central tendency

1. Mean

The value mean for Universiti Canggih is 40.93 and Universiti Muhibah is 57.03. Which we can conclude that students from Universiti Canggih gets the score of 75% above in Online Computer Skills assesment while answering it quicker than Universiti Muhibah.

2. Mode

Mode for Universiti Canggih is 9 with two repetition for each mode while Universiti Muhibah is 4 with three repetition for each mode.

3. Median

The median for Universiti Canggih is 40.5 which is near to its Q1 while Universiti Muhibah is 54.5 which is at center of the box

b) Spread / dispersion

1. Range

the range for Universiti Muhibah is 42 while Universiti Canggih has a range of 50. We can see that Universiti Muhibah has slightly smaller range.

2. Variance

We can see the variance for Universiti Muhibah is 86.03 whereas the Universiti Canggih has a variance of 128.54 meaning that Universiti Canggih has a bigger spread / dispersion between numbers of the dataset.

3. Standard Deviation

We could see that Standard deviation for Universiti Muhibah is 9.28 and for Universiti Canggih is 11.33. Shows that Universiti Canggih has bigger dispersion of dataset relative to its mean.

C.) Shape

1. Skewness

Since the skewness for Universiti Canggih is smaller tan 0 [Skewness < 0]. Its distribution is pointing in the negative direction, while Universiti Muhibah is in the positive direction because the skewness is greater than 0 [Skewness > 0]

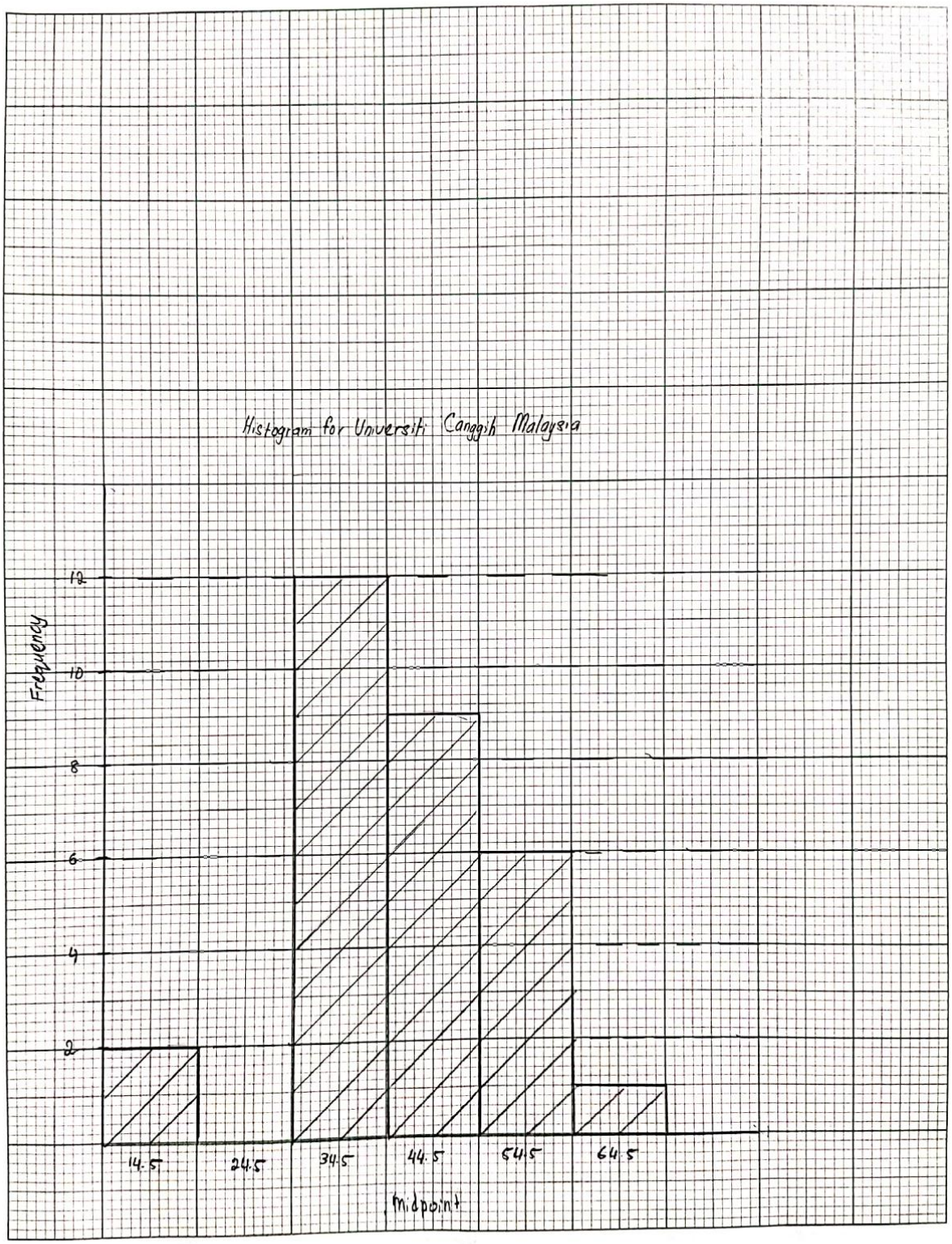
2. Kurtosis

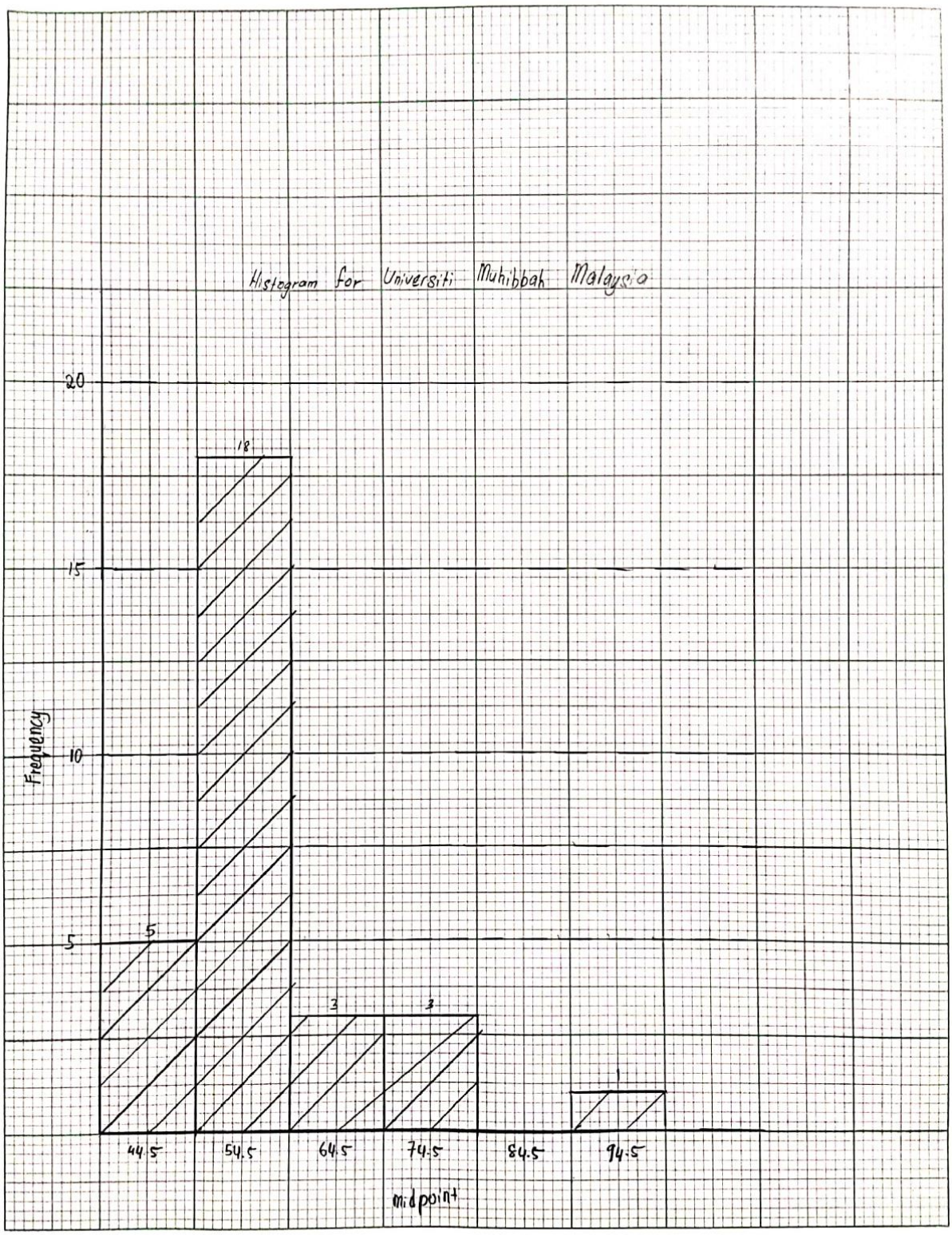
Universiti Canggih has kurtosis of 3.64 and Universiti Muhibah is at 6.36. Both of them are Leptokurtic because the kurtosis is greater than 3.

4.5) Apart from the summary that you produced in (4.1)

a) Suggest a graphical format that can be used for both datasets.
For this datasets, i suggest the graphical format of Histogram.

b) Represent each datasets in graphical format.


Histogram for Universiti Canggih Malaysia
Frequency
12
10
8
6
4
2
14.5
24.5
34.5
44.5
54.5
64.5
Midpoint


Histogram for Universiti Muhibbah Malaysia
20
15
10
5
Frequency
18
5
3
3
1
44.5
54.5
64.5
74.5
84.5
94.5
midpoint

c) Comment on the differences between the two datasets.

Based on the graphical dataset above, the shape for Universiti Canggih is bell shaped and it has a normal distribution. Meanwhile, the shape for Universiti Muhibah is right skewed and it has a assymetrical distribution.

4.6) Do the answering speed of Universiti Canggih look significantly different from Universiti Muhibah?

Yes, the answering speed from Universiti Canggih look significantly different from students from Universiti Muhibah. This is because the value of mean for Universiti Canggih is smaller than Universiti Muhibah which means the average time taken from Universiti Canggih students who managed to get 75% and above in the assesment is 40.93. Other than that, since the data value for Universiti Canggih starts from 10 to 60. That means many students finish getting the score quickly than students from Universiti Muhibah.