SECI 2143 - 01

PROBABILITY & STATISTICAL DATA ANALYSIS

ASSIGNMENT 1

TO : DR AZURAH A SAMAH

FROM :

FARAH NABILAH BINTI MAJMUDIN	A21EC0023
MAATHUREE A/P VEERABALAN	A21EC0051
MUHAMMAD SAIFUDDIN BIN ISMAIL	A21EC0093
NUR SYAFIKA BINTI MOHD SALMIZI	A21EC0115

Question 1

1.1) Descriptive ~~data~~ statistics summarises the characteristics of a data set by collecting, summarizing and describing data.
Inferential statistics can be used for testing hypothesis or making decisions.

1.2) The population used in this study is from students in Malaysia.
A specific sample we used are Computer Science students.

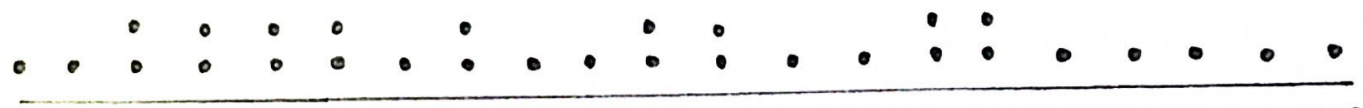
1.3) Regarding the study, we did a survey to discover the internet usage among students. We collect the data by using google form and answered by the students. By specific questions given is google form, we can conclude our hypothesis and relate it with our main goals. Considering the valid data only, we summarize and interpret the data into graphically or numerically for better comparison and research purpose.

1.4) Nominal - gender (female / male)
Ordinal - internet speed (bad, moderate, great)
Interval - CGPA Students
Ratio - age

1.5 https://forms.gle/E7aENjxdWamSExQJ7

Q2

2.1 a)



2.1 b)

stem	leaf
1	0 5
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

key : 1|0 = 10 second

2.2) second fastest highest = 15 second

$y[k] = 15$

$$\text{percentile of value 15} = \frac{\text{number of value less than 15}}{\text{total number of value}} (100)$$

$$= \frac{1}{21} (100)$$

$$= 4.76$$

$$\approx 4$$

The value of 15 is the 4th percentile.

Q3 3.1)

class	0-10	10-20	20-30	30-40	40-50	50-60	
mid value	5	15	25	35	45	55	
frequency	7	10	23	51	6	3	
cumulative frequency	7	17	40	91	97	100	
mid value x freq	35	150	575	1785	270	165	$\Sigma f_i x_i = 2980$

percentile more than 40

3.2 d) $P_? = 41$

$$\text{Percentile} = X + \frac{0.5(Y)}{100} \times 100$$

$$= 91 + \frac{0.5(1)}{100} \times 100 \qquad \text{Percentile} = 91$$

$$= 91.50$$

$$\approx 91$$

3.3 → mean, $\bar{x} = \frac{\Sigma fx}{n} = \frac{2980}{100} = 29.80$

→ median

$L = 30$

$N = 100$

$cf_p = 40$

$w = 10$

$f_{med} = 51$

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

$$= 30 + \left(\frac{\frac{100}{2} - 40}{51}\right)(10)$$

$$= 31.96$$

→ mode

$I = 30$

$h = 10$

$f_1 = 51$

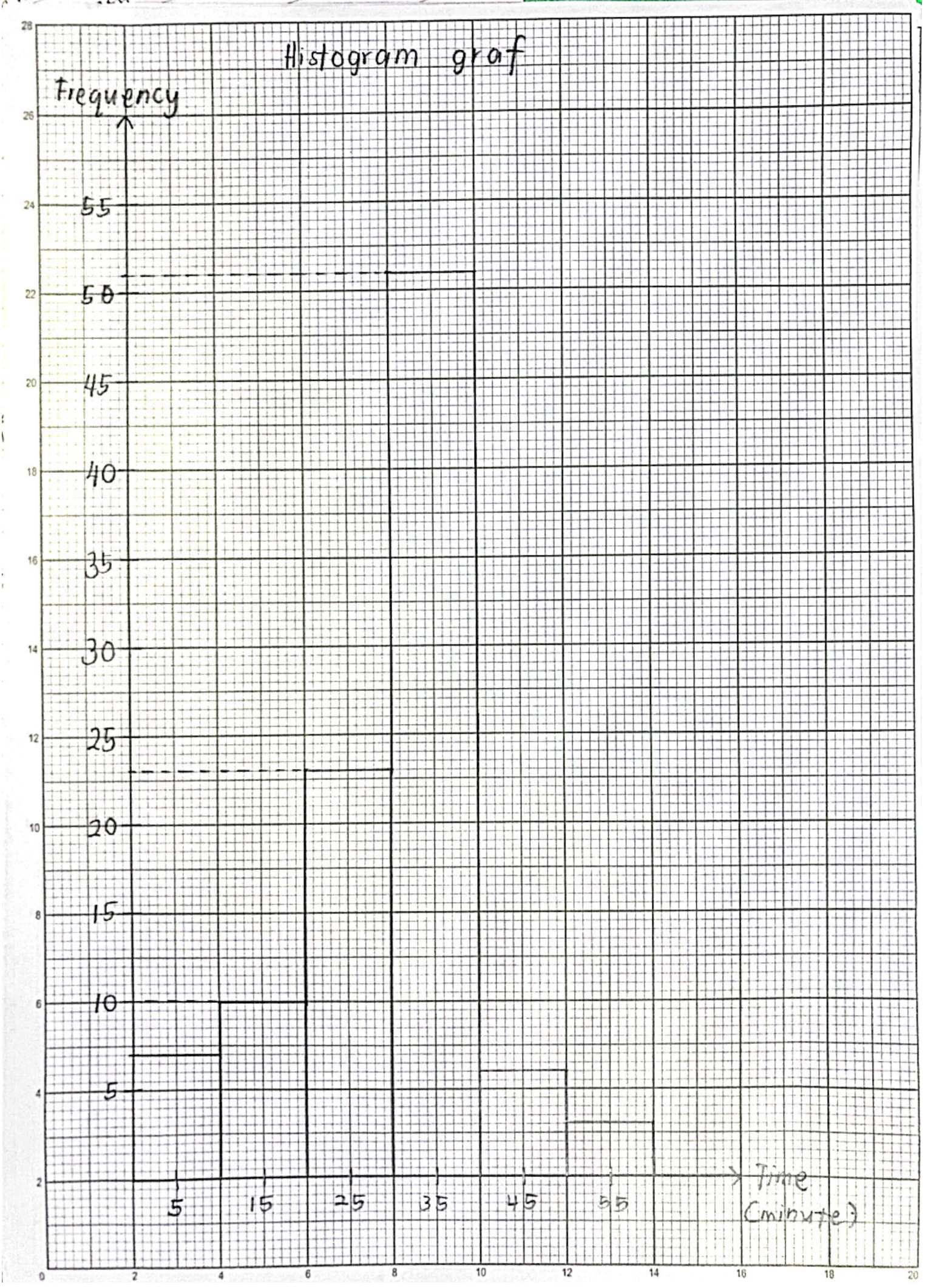
$f_0 = 23$

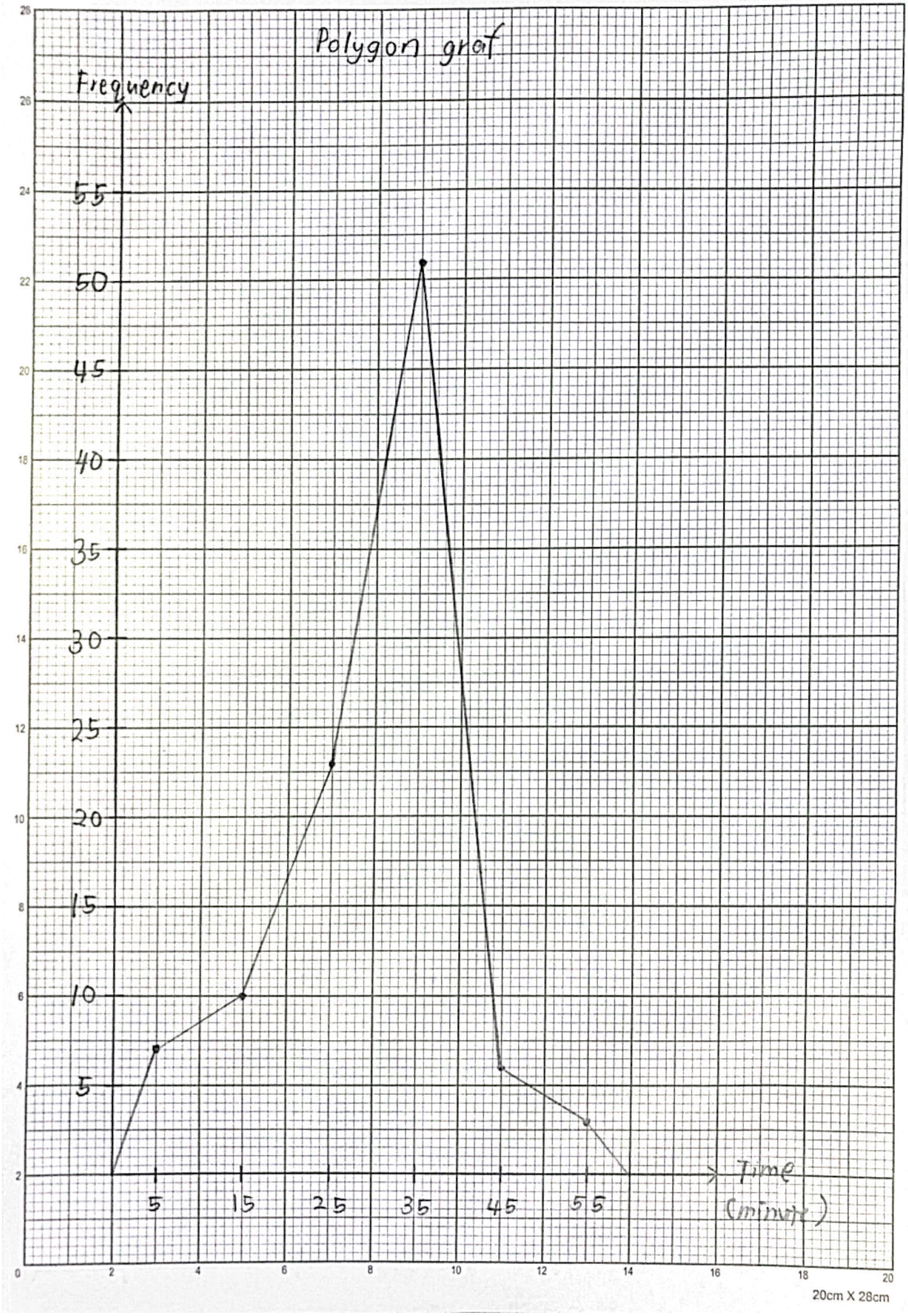
$f_2 = 6$

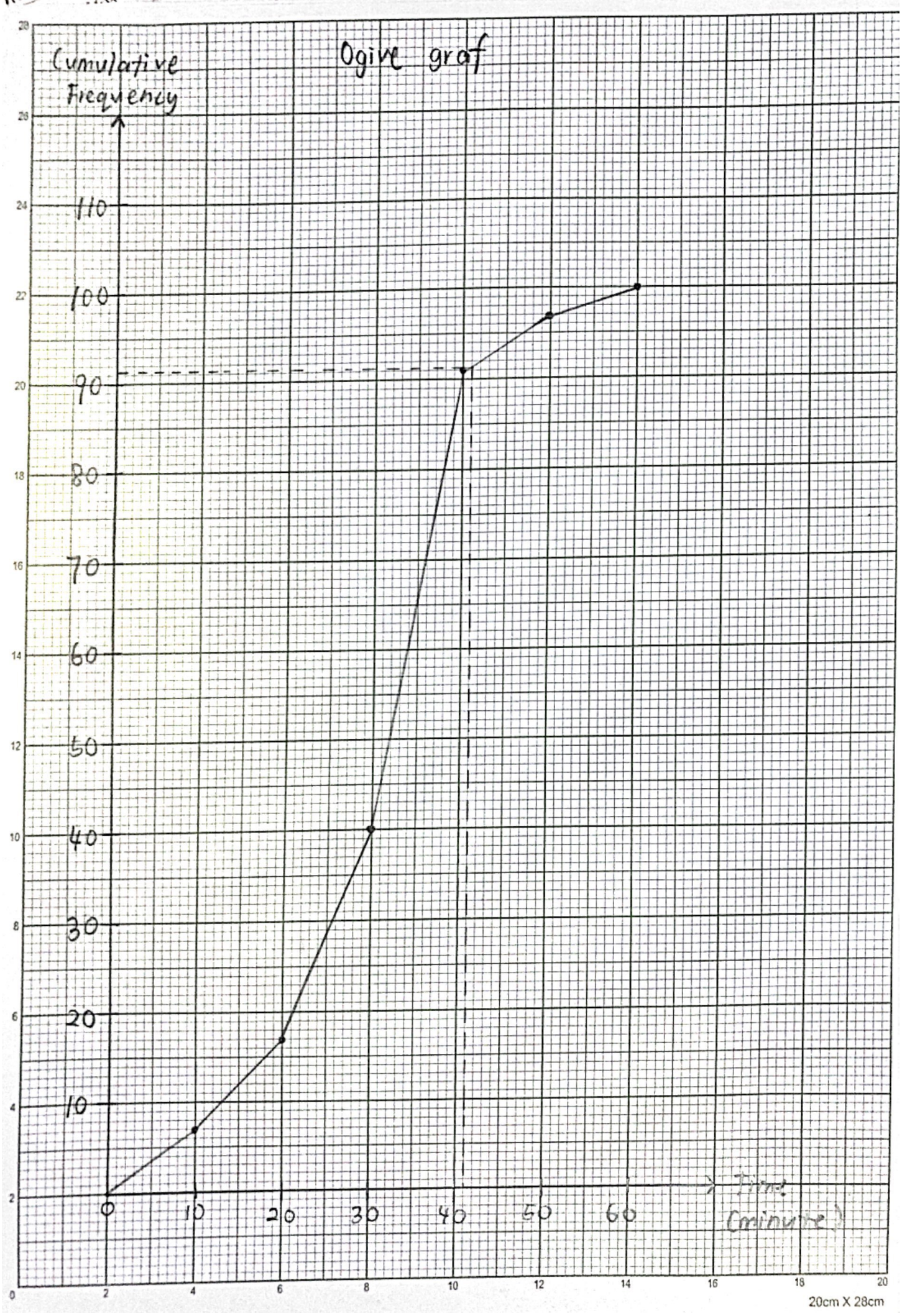
$$\text{mode} = I + h\left[(f_1 - f_0) \div (2f_1 - f_0 - f_2)\right]$$

$$= 30 + 10\left[(51 - 23) \div (2(51) - 23 - 6)\right]$$

$$= 33.84$$







4.1) a) Data set 1 = R = 60 - 10
= 50

Data Set 2 = R = 90 - 48
= 42

b) Data set 1 = $\frac{10+15+32(2)+33(2)+34(2)+35(2)+37+38(2)+39+40+41(2)+44(2)+45+47+48+2)+52(2)+53+55+56+60}{30}$

$= \frac{1228}{30} = 40.93$

Data set 2 = $\frac{48(2)+49(3)+50(3)+51+52(2)+53+54(3)+55(2)+56+57(3)+59(2)+60+63+65+70+72+73+}{30}$

$= \frac{1711}{30} = 57.03$

c) Data Set 1 = 10, 15, 32, 32, 33, 34, 34, 35, 37, 38, 39, 40, 41, 44, 44, 45, 47, 48, 52, 52, 53, 53, 55, 56, 57, 60

n = 30 - even number

= (40 + 41) ÷ 2 = 40.5

Data Set 2 = 48, 48, 49, 49, 49, 50, 50, 50, 51, 52, 52, 53, 54, 54, 55, 56, 57, 57, 59, 59, 60, 63, 65, 70, 72, 73, 90

n = 30 - even number

= (54 + 55) ÷ 2 = 54.5

d) Data set 1 :

32, 33, 34, 35, 38, 41, 44, 48, 52 - occur two times

- the modes are 32, 33, 34, 35, 38, 41, 44, 48, 52

Data set 2 = 49, 50, 54, 57 - occur three times

the mode are 49, 50, 54, 59

d) Data set 1 : $N = 30$ $Q_1 = P = 25$ $k = 8$

$i = \frac{25}{100}(30)$ $Y(8) = 34$

$= 7.5$ $Q_1 = 34$

$= 8$

Data set 2 : $N = 30$ $Q_1 = P = 25$

$k = 8$

$i = \frac{25}{100}(30)$ $Y(8) = 50$

$= 7.5$ $Q_1 = 50$

$= 8$

e) Data set 1 :

$N = 30$ $Q_3 = P = 75$

$i = \frac{75}{100}(30)$ $k = 23$

$Y(23) = 48$

$= 22.5$ $Q_3 = 48$

$= 23$

Data set 2 :

$N = 30$, $Q_3 = P = 75$ $k = 23$

$i = \frac{75}{100}(30)$ $Y(23) = 59$

$= 22.5$ $Q_3 = 59$

$= 23$

i) Data set 1 :

$n = 30, \bar{x} = 40.93, s = 11.34$

$$= \frac{(110 - 40.93)^3 + (15 - 40.93)^3 \ldots + (60 - 40.93)^3}{(30-1)(11.34)^3}$$

$$= -0.66$$

Data set 2 :

$n = 30 \quad \bar{x} = 57.03, \quad s = 9.28$

$$= \frac{(48 - 57.03)^3 + (48 - 57.03) \ldots (60 - 57.03)^3}{(30-1)(9.28)^3}$$

$$= 7.56$$

ii) Data set 1

$n = 30 \quad \bar{x} = 40.93 \quad s = 11.34$

$$= \frac{(110 - 40.93)^4 + (15 - 40.93)^4 \ldots (60 - 40.93)^4}{(30-1)(11.34)^4}$$

$$= 3.64$$

Data set 2 :

$n = 30 \quad \bar{x} = 57.03 \quad s = 9.28$

$$= \frac{[(48 - 57.03)^4 + (48 - 57.03)^4 \ldots (60 - 57.03)^4]}{(30-1)(9.28)^4}$$

$$= 6.37$$

f) Data set 1 :

$IQR = Q_3 - Q_1$

$= 48 - 34$

$= 14$

Data set 2 =

$IQR = Q_3 - Q_1$

$= 59 - 50$

$= 9$

g) Data set 1 :

$n = 30 \quad \bar{x} = 40.93$

$$s^2 = \frac{((10 - 40.93)^2 + (15 - 40.93)^2 + (32 - 40.93)^2 \ldots (60 - 40.93)^2)}{30 - 1}$$

$= 128.55$

Data set 2 :

$n = 30 \quad \bar{x} = 57.03$

$$s^2 = \frac{((48 - 57.03)^2 + 48 - 57.03)^2 \ldots + (80 - 57.03)^2)}{30 - 1}$$

$= 86.03$

h) Data set 1 : $s^2 = 128.55$

$s = \sqrt{128.55}$

$= 11.34$

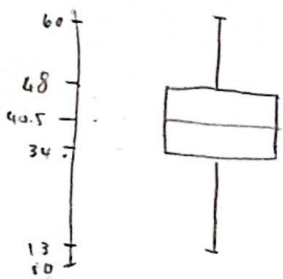
Data set 2 : $s^2 = 86.03$

$s = \sqrt{86.03}$

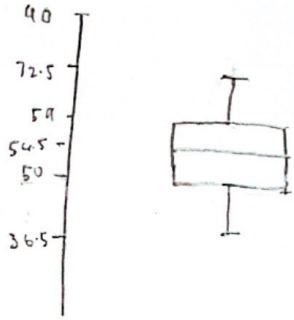
$= 9.28$

4.2) There are outliner for both sets. The outliner for set 1 is 10 as it Dutside of the adjacent value region-while in data set 2, the outliner are 33 and 80 as they fall outside of adjacent value region.

4.3.) Data set 1



Data set 2



4.4 (a) The measurement of central tendency include the mean, mode and median of the datasets.

The median value of first datasets is 40.5 which is smaller than median of datasets 2, 54.5.

The upper part of the box of the median datasets 1 is longer than the lower part.
It is negatively skewed.
While, the upper part of the box of the median datasets 2 is the same as the lower part.
Therfore, there is no skew (normal) disttribution).

The mean of datasets 1 was smaller than datasets 2, 40.93 compared to 57.03.

Both the datasets are multimodal.

(b) The measurement of dispersion include the range, variance and standard deviation of the datasets.

The value of range for datasets 1 is bigger than datasets 2, 50 compared to 42.

The variance for datasets 1 also bigger than datasets 2 which is 128.55 compared to 86.03.

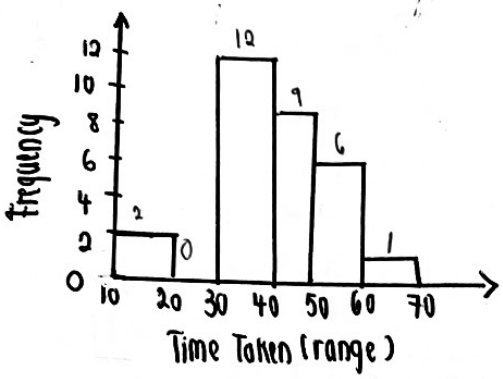
The standard deviation of datasets 1 is 11.34 and it is bigger than standard deviation of datasets 2, 9.28.

(c) The first datasets of boxplot is negatively skewed (skewed left) while the second datasets is symmetric (no skew).

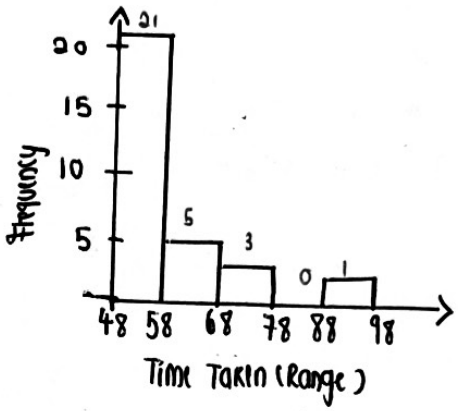
The box-plots datasets 1 interquartile range (IQR) is bigger than box-plot database 2 which is 14 and 9.

4.5 (a) Histogram

(b) First datasets :



Second datasets :



(c) The right skewed shape is second datasets because it has long "tail" on the right side of the distribution.

The histogram shows that students from datasets 1 speed is more than 50 minutes while students from datasets 2 take less than 60 minutes. For datasets 1, the answering speed is positive skew while the answering speed for datasets 2 is symmetric.

4.6 Yes, the answering speed between both universities are different.

Based on the histogram, the answering speed of students from University Canggih is more faster than students from Universiti Muhibbah.

Most of the students from Universiti Muhibbah takes more than 90 minutes time to answer the assessment while most of the students from Universiti Canggih takes less than 90 minutes to answer the assessment.