TOPC 1: STATISTICS

Statistics is a branch of Mathematics that deals with observation, collection, recording, analysis and interpretation of data gathered from a field of study.

Descriptive Statistics: This deals with collection of data such as population, rainfall received etc.

Data 0rganisation:
Data may be thought of as being results of an observation or experiment. It can be referred to as raw facts or un processed information.

Some basic terms:

1. **Population:** Any set of items under study (consideration) which we define by some shared characteristics. A population does not necessarily mean people.
   Examples
   - A population of papers in the book
   - A population of road accidents etc.

   Population can be finite or infinite. Finite population is when the number of population is known otherwise infinite

2. **Sample:** This is when the observation is made from the subset of the population
3. **Variable:** This is the quantity that is observed in experiment and it varies from each member of the population
   Types of Data
   Any data collected can be classified into;
   i) **Qualitative data:** This data measures attributes that can not be quantified such as sex, color, etc
   ii) **Quantitave data:** This is data which can be quantified. It can be represented in numerical form. Eg height, mass, distance,etc

   Quantitave data can be further subdivided into two;

a). Discrete data: This is data which takes on countable descriptive values ie 0,1,2, 3,....

E.g

- ✓ Number of students in class
- ✓ Number of cars in a car yard

b) **Continuous Data:** This is data that takes on values in a given range or Interval ie $(a,b) \Rightarrow a < x < b$.

Example: Age of students below 10 years, 0<x<10

Methods of data Presentation

a) By use of frequency tables.

b) By use of pie charts, bar charts.

c) By use of graphs e.g Ogives, frequency polygon, histograms etc

Use of pie charts

Example 1:

Given the data below;

Method of travel	Number of pupils
Bus	12
Car	2
Bicycle	5
Walking	9

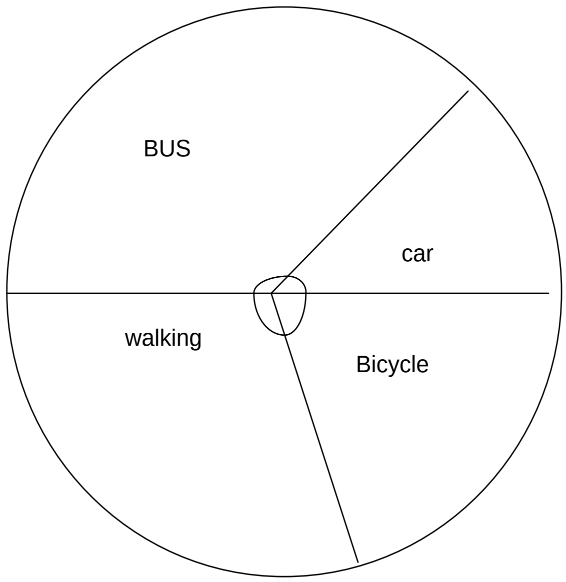
a) Calculate to the nearest degree the sector angles of the pie chart.

b) Draw a pie chart using a circle of radius of 5cm, labeling each sector with a method of travel it represents.

Solution

Bus $\frac{12}{28} \times 360^0 = 154^0$ *Bicycle* $\frac{5}{28} \times 360^0 = 64^0$ *car* $\frac{2}{28} \times 360^0 = 26^0$ *walking* $\frac{9}{28} \times 360^0 = 116^0$

A pie chart representing the method of travel



Frequency tables:

These are used to show frequency of different scores. They help in organizing of data.

Parameters for ungrouped data

1.1 UNGROUPED DATA

Measures central tendancy for ungrouped data.

The most frequently encountered summary measures in statistics are measures of central tenadancy. They tend to locate the central value.

a) Mean, $\bar{x} = \frac{sum\ of\ values}{total\ number\ of\ values} = \frac{\sum x}{n}$, where x is for data values.(given set of values)

   If the frequency table is given then

   $\bar{x} = \frac{\sum fx}{\sum f}$

   When the working mean/ assumed mean, A, is given then

   Where $d = x - A$

   $\bar{x} = A + \frac{\sum d}{n}$,For a given set of values

$\bar{x} = A + \frac{\sum fd}{\sum f}$, For a given frequency table

b) The mode: It is that observation that occurs more frequently.

c) Median: It is the middle value of the distribution that is obtained after the values have been arranged in order.
If the order of the distribution is odd, then the median is the value in the position $\left(\frac{n+1}{2}\right)^{th}$. If the order of the distribution is even then there are two middle values whose positions are $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+2}{2}\right)^{th}$. These are positions of two middle values. The median is the average of the values in those positions.

Example 2:

The time in seconds for phone calls made by 12 customers at public telephone booth were recorded as follows:
101, 132, 101, 91, 89, 122, 115, 106, 109, 112, 105, 106.
Find the:

i) Median time
ii) Mean time
iii) mode

soln

i) ~~89,91,101,105,106~~,106, 109,1~~10, 112,115,122,132~~

Median= $\frac{106+109}{2} = 107.5$

ii) Mean= $\frac{\sum x}{n}$,

$\sum x = 89 + 91 + 101 + 105 + 106 + 106 + 109 + 110 + 112 + 115 + 122 + 132 = 1298$

Mean= $\frac{1298}{12} = 108.1667$

OR

A frequency table can be used

X	f	fx
89	1	89
91	1	91
101	1	101
105	1	105
106	2	112
109	1	109
110	1	110
112	1	112
115	1	115
122	1	122
132	1	132
	$\sum f = 12$	$\sum fx = 1298$

Mean, $\bar{x} = \frac{\sum fx}{\sum f} = \frac{1298}{12} = 108.1667$

Measures of variation of ungrouped data

While measures of central tendancy locate the centre of the distribution measures of dispersion permit an assessment to be made of the extent to which data spreads. How a variable spreads away from the mean

1. Range: It is the difference between the largest value and the smallest value
2. Variance= $\frac{\sum (x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$, when the frequency distribution is given then

   Variance= $\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$

3. Standard deviation= $\sqrt{\text{var}iance}$

4. **MEAN DEVIATION**

This measures the deviation of each score or variable from the mean without regard to the sign.

$$mean\ deviation = \frac{\sum|x-\bar{x}|}{n}$$

Note:

Properties of summation $(\sum\ )$

1. $\sum(x+y+z)=\sum x+\sum y+\sum z$
2. $\sum Cx = C\sum x$ **, where C is a constant**
3. $\sum_{1}^{n} C = nC$ **,Where C is a constant.(summation of a constant n times=nC)**

Proof:

1. $\bar{x} = \frac{\sum x}{n} = A + \frac{\sum d}{n}$

From d=x-A

Multiply by $\sum$

$$\sum d = \sum x - \sum A$$
$$\sum d + \sum A = \sum x$$
$$but, \bar{x} = \frac{\sum x}{n}$$
$$\bar{x} = \frac{\sum d + \sum A}{n} = \frac{\sum d}{n} + \frac{\sum A}{n} = \frac{\sum d}{n} + \frac{nA}{n},$$

Since A is a constant $\sum A = nA$

$$\bar{x} = A + \frac{\sum d}{n}$$

2. Show that $\bar{x} = \frac{\sum fx}{\sum f} = A + \frac{\sum fd}{\sum f}$

From d=x-A

Multiply by $\sum f$

$$\sum fd = \sum fx - \sum fA$$

$$\sum fx = \sum fd + \sum fA$$

$$but, \bar{x} = \frac{\sum fx}{\sum f}$$

$$\bar{x} = \frac{\sum fd + \sum fA}{\sum f}$$

Since A is a constant

$$\frac{\sum fA}{\sum f} = \frac{A\sum f}{\sum f}$$

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

3. $Variance = \dfrac{\sum (x-\bar{x})^2}{n} = \dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2$

$$(x-\bar{x})^2 = x^2 - 2x\bar{x} + \bar{x}^2$$

$$\sum (x-\bar{x})^2 = \sum x^2 - 2\bar{x}\sum x + \sum \bar{x}^2$$

$$\frac{\sum (x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - 2\bar{x}\frac{\sum x}{n} + \frac{\sum \bar{x}^2}{n}$$

$$= \frac{\sum x^2}{n} - 2\bar{x}^2 + \frac{n\bar{x}^2}{n}$$

$$= \frac{\sum x^2}{n} - \bar{x}^2$$

$$= \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$$

Activity

4. **Show that** $\dfrac{\sum f(x-\bar{x})^2}{\sum f} = \dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f}\right)^2$

Example 3:

Given the data below;

3, 6, 7, 14, 14. Find the variance and Standard deviation.

Soln:

$$mean, \bar{x} = \frac{3+6+7+10+14}{5} = 8$$

$$\sum x^2 = 3^2 + 6^2 + 7^2 + 10^2 + 14^2 = 390$$

Standard deviation= $\sqrt{\text{var}iance}$

$$\text{var}iance = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{390}{5} - 8^2 = 14$$

Standard deviation= $\sqrt{14} = 3.742$

Example: Given that $\sum f = 20, \sum fx^2 = 16143, \sum fx = 563$

Determine the mean and Variance

$$mean, \bar{x} = \frac{\sum fx}{\sum f} = \frac{563}{20} == 28.15$$

$$\text{var}iance = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{16143}{20} - 28.15^2 = 14.7275$$

Quartiles: A Quartile is a value that divides the given data into four equal parts

For ungrouped data;

Upper quartile, q_3, is the value that divides data on the right of the median into two equal parts.

Lower quartile, q_1, is the value that divides values on the left of the median into two equal parts

The inter-quartile range= $q_3 - q_1$

Semi – inter-quartile range= $\frac{q_3 - q_1}{2}$

Example 4:

Given the data below

1,4,6,10,9,5,6,5,2,2,3,7,4,5,6,4,4

Determine: the mean, mode and the median

Solution:

We can use a frequency table:

X	f	fx	Cf

1	1	1	1
2	2	4	3
3	1	3	4
4	4	16	8
5	3	15	11
6	3	18	14
7	1	7	15
9	1	9	16
10	1	10	17
	$\sum f = 17$	$\sum fx = 83$	

Mode=4

$$\text{Mean}=\frac{\sum fx}{\sum f}=\frac{83}{17}=4.8823$$

Median: Since the data is odd the position of the median is given by

$$\left(\frac{n+1}{2}\right)^{th}=\left(\frac{17+1}{2}\right)^{th}=9^{th}\text{ postion}$$

The value in the 9^{th} position is 5

The median=5

Example 5:

Given the data below:

marks	10	11	12	13	14	15	16	17	18	19	20
F	1	2	2	2	2	4	2	1	2	1	1

Find:

a) The mean
b) Standard deviation
c) The inter-quartile range

Solution

Refer to the table beow

a) $\bar{x} = \frac{\sum fx}{\sum f} = \frac{294}{20} = 14.7$

b) Standard deviation=

$$\sqrt{variance} = \sqrt{\left(\frac{\sum fx^2}{\sum f} - \bar{x}\right)} = \sqrt{\left(\frac{4470}{20} - 14.7^2\right)} = 2.722$$

x	f	fx	x²	fx²
10	1	10	100	100
11	2	22	121	242
12	2	24	144	288
13	2	26	169	338
14	2	28	196	392
15	4	60	225	900
16	2	32	256	512
17	1	17	289	289
18	2	36	324	648
19	1	19	361	361
20	1	20	400	400
	$\sum f = 20$	$\sum fx = 294$		$\sum fx^2 = 4470$

c) 10,11,11,12,12,13,13,14,14,15,15,15,1516,16,17,18,18,19,20

q_1 *lowerquartile* q_2 *median* q_3 *upperquartile*

Since the data is even we get two middle values as shown above

$q_1 = \frac{12+13}{2} = 12.5$ $\quad$ $q_3 = \frac{16+17}{2} = 16.5$

Inter-quartile range=16.5-12.5=4

Example 6:

The weight, x, grams of 30 chocolate buns are summarized as shown below;

$\sum(x-k) = 315$, $\sum(x-k)^2$ =4022, where k is a constant. The mean weight of the buns is 50.5g. Find the;

i) Value of k

ii) Standard deviation

Solution

i) Mean= $\frac{\sum x}{n} = 50.5$

$\sum x = 50.5n = 50.5 \times 30 = 1515$

$\sum(x-k) = \sum x - \sum k = 315$, since k is a constant, $\sum k = nk$ =30k

1515-30k=315

K=40

ii) $\sum(x-k)^2$ =4022

$$\sum(x-40)^2 = 4022$$

$$\sum(x^2 - 80x + 1600) = 4022$$

$$\sum x^2 - 80\sum x + \sum 1600 = 4022$$

$$\sum x^2 - 80 \times 1515 + 1600 \times 30 = 4022$$

$$\sum x^2 = 77222$$

$$Variance = \frac{77222}{30} - 50.5^2 = 23.8167$$

Activity:

1. A class performed an experiment to estimate the diameter of a circular object, a sample of 5 scholars had the following results 3.12, 3.16, 2.94, 3.33 and 3.0

a) Determine the sample mean
b) Standard deviation

2. Given that the $n = 20$, $\sum x^2 = 1647$, $\sum x = 56$. Find the mean and variance

3. a) A set of values has **m** zeros and **n** ones. Find the mean of the se of data, hence show that the standard deviation$=\dfrac{\sqrt{mn}}{m+n}$

b). Given the data below;

x	1	2	3	4	5
f	a	11	b	8	9

Given that the mean is 2.7 and $\sum f = 50$, find a and b.

5. The numbers a, b, 8, 5 and 7 have a mean of 6 and variance of 2. Find the values of a and b if a>b

1.2. GROUPED DATA (CONTINUOUS DATA)

Continuous data cannot show an actual value but can only be given with a certain range:

Ways of grouping data

a)

Length(mm)	f
27 – 31	4
32 – 36	11
37 - 46	12

Class widths are 5,5,10

Class boundaries are 26.5,31.5,36.5 and 46.5

b)

Length(mm)	f

0-	9
3-	12
9-	3
14 - 17	2

Class widths are 3, 6, 5 and 3

Class boundaries are; 0,3,6,9 and 14

c).

Mass(g)	F
-50	8
-100	10
-150	16
-250	6

The class boundaries are 0, 50, 100, 150, and 200

The class widths are 50, 50, 50, and 100

d).

Speed (km/hr)	f
20-30	2
30-40	7
40-50	3
50- 70	1

Class boundaries are 20,30,40,50 and 70

The class widths are 10, 10, 10, and 20

e).

Time (s)	f
20 - <30	2
30 - <40	3
40 - <60	1

The class boundaries are 20, 30, 40 and 60

The class widths are 10, 10 and 20

Note:

Class width = (upper class boundary) - (lower class boundary)

Measures of central tendency for ungrouped data

a) Mean, $\bar{x} = \frac{\sum fx}{\sum f}$, When the working mean is given $\bar{x} = A + \frac{\sum fd}{\sum f}$

b) Mode

Mode$= L_1 + \left(\frac{d_1}{d_1 + d_2}\right)C$

Where L_1=lower class boundary of the modal class

For data with equal class widths

d_1= (modal frequency) – (frequency before that of the modal class)

d_2 =(Modal frequency) – (frequency after that of the modal class)

C=Class width

For data with unequal class widths

$d_1 = \left(\begin{matrix}modal\ frequeny \\ density\end{matrix}\right) - \left(\begin{matrix}frequency\ density\ before\ that\ of the \\ modal\ class\end{matrix}\right)$

$d_2 = \left(\begin{matrix}\boldsymbol{modal\ frequency} \\ \boldsymbol{density}\end{matrix}\right)$-$\left(\begin{matrix}frequency\ density\ after\ that\ of the \\ modal\ class\end{matrix}\right)$

C=class of the modal class

Note: $frequency\ density = \frac{frequency}{class\ width}$

c) Median

Median= $L_m + \left(\dfrac{\frac{N}{2} - CF_b}{f_m} \right) C$

Where N= $\sum f$

$\frac{N}{2}$ = position of the median

L_m =lower class boundary of the median class

f_m = Frequency of the median class

CF_b = Cumulative frequency before that of the median class

C= class width of the median class

Measures of Variation fo grouped data

Variance = $\dfrac{\sum f(x-\bar{x})^2}{\sum f} = \dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f} \right)^2$

Standard deviation= $\sqrt{Variance}$

Quatiles

a) **Lower quartile;** $q_1 = L_{q_1} + \left(\dfrac{\frac{N}{4} - CF_b}{f_{q_1}} \right) C$

Where;

L_q =lower class boundary of the lower quartile class

N= total frequency

CF_b = Cumulative frequency before that of the lower quartile class

f_q = frequency of the lower quartile class

C= class width of the lower quartile class.

$(\frac{N}{4})^{th}$ = position of the lower quartile class.

b) **Upper quartile;** $q_3 = L_q + \left(\dfrac{\frac{3N}{4} - CF_b}{f_q} \right) C$

Where L_q =lower class boundary of the upper quartile class

N= total frequency

CF_b = Cumulative frequency before that of the upper quartile class

f_q = frequency of the lower quartile class

C= class width of the upper quartile class.

$(\frac{3N}{4})^{th}$ = position of the upper quartile class.

Deciles: Deciles are denoted by $D_1, D_2, \ldots\ldots\ldots, D_9$. A decile is a value that divides the given data into 10 equal parts. The k[th] decile, denoted by D_k of grouped data is given by

$$D_k = L_k + \left(\frac{\frac{k}{10}N - CF_b}{f_k} \right) C$$

Where

L_k = Lower class boundary of the k[th] decile class

CF_b = cumulative frequency before that of the k[th] decile class

f_k = frequency of this decile class

Percentiles: percentiles denoted by $P_1, P_2, P_3, \ldots\ldots. P_{99}$. A percentile is a value that divides the given data into 100 equal parts. The k^{th} percentile denoted by

$$P_k = L_k + \left(\frac{\frac{k}{100}N - CF_b}{f_k} \right) C$$

Where L_k = Lower class boundary of the k[th] percentile class

CF_b = cumulative frequency before that of the k[th] percentile class

f_k = frequency of this percentile class

GRAPHS

Histogram: A histogram is a plot of frequency against class boundaries (for equal class widths)

OR A plot of frequency density against class boundaries for unequal class widths

A histogram can be used to find:

i) Mode
ii) median

Ogive/cumulative frequency curve: It is a plot of cumulative frequency against class boundaries. An Ogive can be used to determine:

i) median
ii) quartiles
iii) deciles
iv) percentiles
v) probability

NOTE:

- ✓ **Plotting of class marks for an Ogive should be avoided(wrong)**
- ✓ **The scale on the histogram or Ogive must be uniform and the axes must be labeled**

Example1:

The data below shows the ages of patients who tested positive with COVID 19

Age (years)	Number of patients
40 – 44	9
45 – 49	13
50 – 54	17
55 – 59	10
60 – 64	8
65 – 69	6
70 – 74	2

Calculate the (i) median

(ii) Interquartile range and hence the semi inter quartile range

Solution:

Age	f	CF	Class boundaries
40 – 44	9	9	39.5 – 44.5
45 – 49	13	22	44.5 – 49.5
50 – 54	17	39	49.5 – 54.5
55 – 59	10	49	54.5 – 59.5
60 – 64	8	57	59.5 – 64.5
65 – 69	6	63	64.5 – 69.5
70 - 74	2	65	69.5 – 74.5

i) Position of the median class $=\left(\frac{65}{2}\right)^{th}=32.5^{th}$

Median class is 50 – 54

Median$=49.5+\left(\frac{32.5-22}{17}\right)5=52.5882$

ii) Lower quartile

Position of the lower quartile class$=\left(\frac{65}{4}\right)^{th}=16.25^{th}$

$q_1=44.5+\left(\frac{16.25-9}{13}\right)5=47.2885$

Position of the upper quartile class$=\left(\frac{3(65)}{4}\right)^{th}=(48.75)^{th}$

Upper quartile$=54.5+\left(\frac{48.75-39}{10}\right)5=59.37$

Inter quartile range=upper quartile – lower quartile
=59.37-47.2885=12.0865

Semi interquartile range $=\frac{12.0865}{2}=6.04325$

Example 2:

The table below shows the frequency distribution of marks obtained in a test by a group of senior six students in a certain school

Marks	10-	20-	30-	40-	50-	60-	70-	80 - 90
frequency	18	34	58	42	24	10	6	8

(a) Estimate the mean mark.

(b) Draw the cumulative frequency curve. From your graph, estimate

(i) the median mark

(ii) How many would fail if the pass mark is fixed at 40.

(iii) The range of values within which the middle 50%

of the insect lies.

soln

Marks	f	x	fx	C.f
10-20	18	15	270	**18**
20-30	34	25	850	**52**
30-40	58	35	2030	**110**
40-50	42	45	1890	**152**
50-60	24	55	1320	**176**
60-70	10	65	650	**186**
70-80	6	75	450	**192**
80-90	8	85	680	**200**
	$\sum f$ = **200**		$\sum fx$ = **8140**	

(a) Mean, $\bar{x} = \frac{8140}{200} = 40.7$

(b)(i) median=38

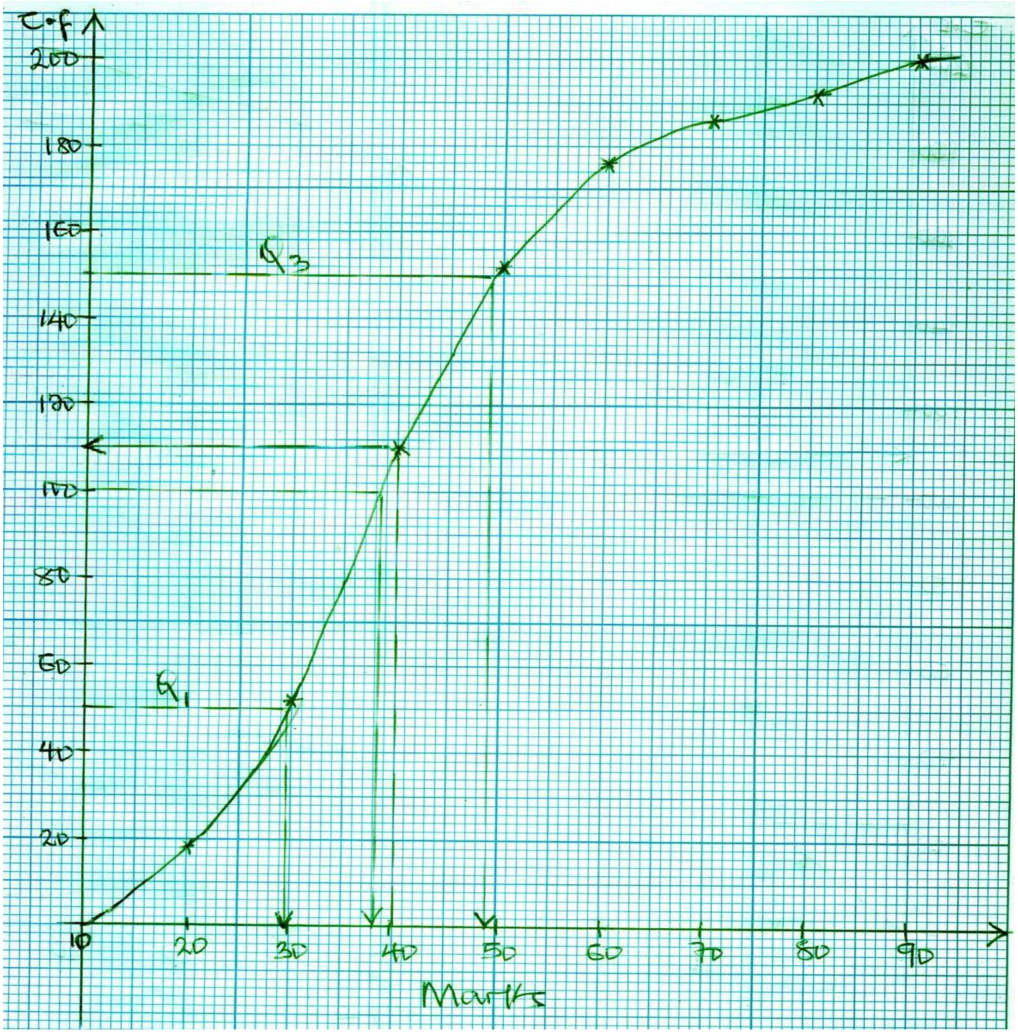
(ii) 110 students

(iii)middle 50%

$P_{75} = \frac{75}{100} \times 200^{th} = 150^{th}$ the value is 49

$P_{25} = \frac{25}{100} \times 200^{th} = 50^{th}$, the value is 30

Range=49-30=19



Example 3:

The frequency distribution below shows the ages of 240 students admitted to a certain university

Ages(years)	Number of students
18-<19	24
19-<20	70
20-<24	76
24-<26	48
26-<30	16
30-<32	6

a) Calculate the mean age of students

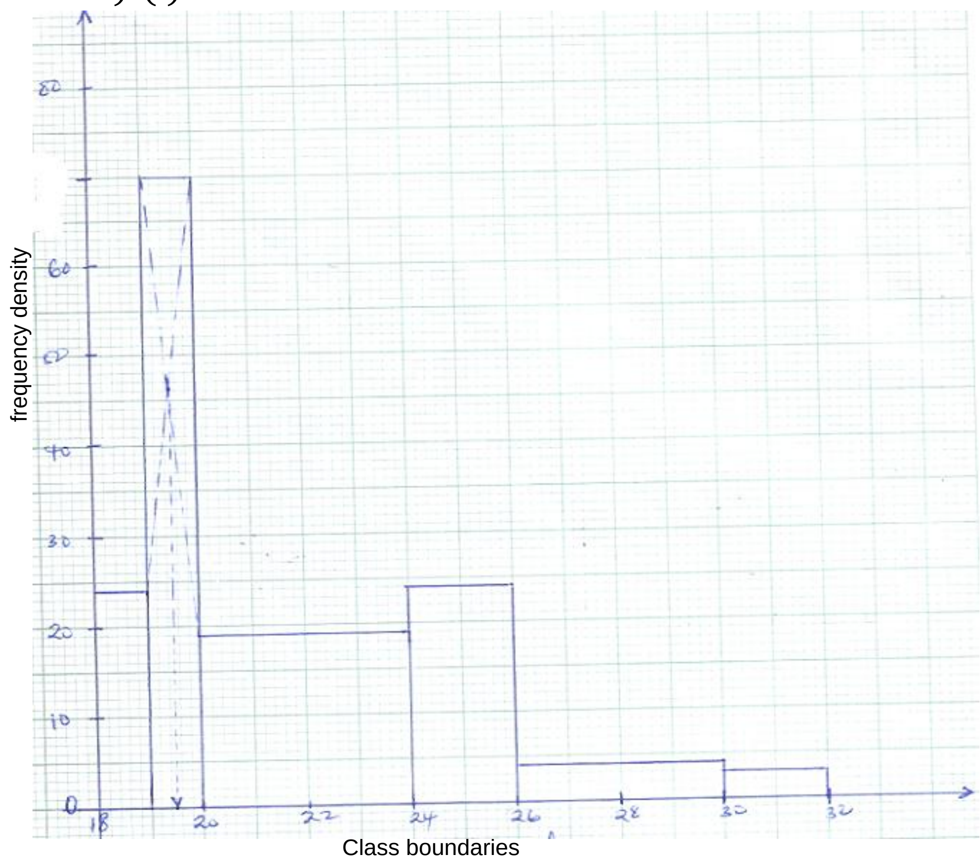
b) (i) Draw the histogram for the data

(ii) Use the histogram to estimate the modal age
Solution:

Age(years)	f	x	fx	Class boundaries	Class width	Frequency density
18-<19	24	18.5	444	18-19	1	24
19-<20	70	19.5	1365	19-20	1	70
20-<24	76	22.0	1672	20-24	4	19
24-<26	48	25.0	1200	24-26	2	24
26-<30	16	28.0	448	26-30	4	4
30-<32	6	31.0	186	30-32	2	3
Total	240		5315			

a) Mean age= $\frac{\sum fx}{\sum f} = \frac{5315}{240} = 22.1448 years$

b) (i)



Assignment:1.1.1

1. The times to the nearest second taken by 100 students to solve a given problem are shown below.

Times(seconds)	Number of Students
30 – 49	10
50 – 64	30
65 – 69	25
70 – 74	20
75 - 99	15

Find the:

(a) Mean

(b) Modal time of the distribution

2. The table below shows the distribution of weights of certain type of mango.

Weight(g)	<10	<15	<30	<45	<55	<65	<80	<95
frequency	4	3	23	54	16	9	7	4

a) Calculate

i) Mean weight

ii) Standard deviation of the distribution

3. The table below shows the distribution of weights of certain type of mango.

Weight(g)	<10	<15	<30	<45	<55	<65	<80	<95
frequency	4	3	23	54	16	9	7	4

b) Calculate

iii) Mean weight

iv) Standard deviation of the distribution

4. The table below represents the masses of donkeys and their respective frequency densities.

Mass(kg)	12-20	20-24	24-30	30-32	32-38	38-48	48-60
Frequency density	1	6	4	8	2	1	0.5

(a) Draw a histogram and use it to estimate the modal mass

(b) Calculate the:

i) Median mass

ii) Number of donkeys with a mass less than 33kg

iii) 60^{th} percentile of the distribution