R-studio

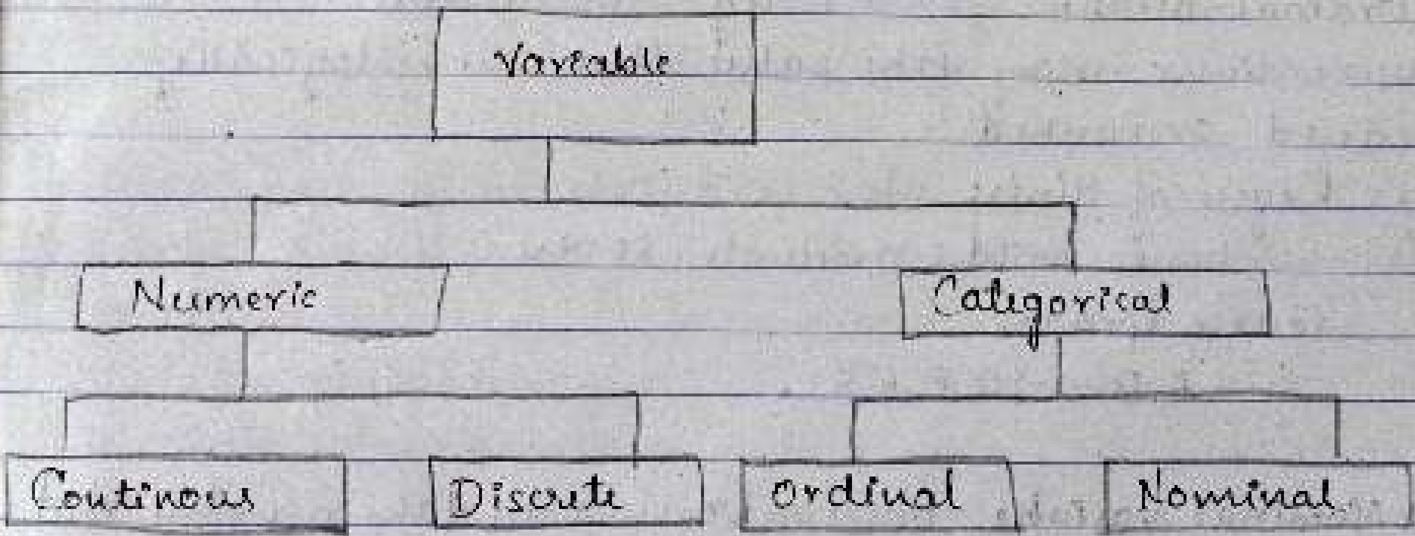
Applied Statistics

A statical enquery has four phases.
1) Collection of data
2) Classification and tabulation of data
3) Analysis of data
4) Interpritation of data

1) Collection of data

Data means information. Data collected expressly for a specific purpose are called primary data
Ex: Data collected by a particular person or organization from the primary source
Data collected and published by one organization and subsequently used by other organization are called as secondary data. The various sources of collection for secondary data are:- newspapers and periodicals

Data will be collected on individuals from the population and termed as variables. This variables are the characteristics of the individuals within the population



* Numeric variables are variables that take numeric measures upon which arithmatic operation can be carried out on the characteristics of individuals. Numeric variables are further classified as:

i) Discrete variable is a quantative variable that will assume a finite or countable set of values.

Ex: No. of students late for the class
No. of children in the family.
SAT scores
No. of crimes reported to police

ii) Continous variable is a quantitative variable that has an infinite no. of values in other words a continous variable can assume any value b/w any two points on the real line

Ex: cholesterol level, height, age

* Categorical variables have values that describe a quality or characteristic of a data unit. Categorical variables may be further described as ordinal & nominal

i) Ordinal variable is a categorical variable Observations may take value that can be logically ordered or ranked.

Ex: Degree of illness
- none, mild, moderate, severe
Course Grades
A, B, C, D, F, E, S

ii) Nominal variable observations can take a value that is not able to be organized in a logical sequence

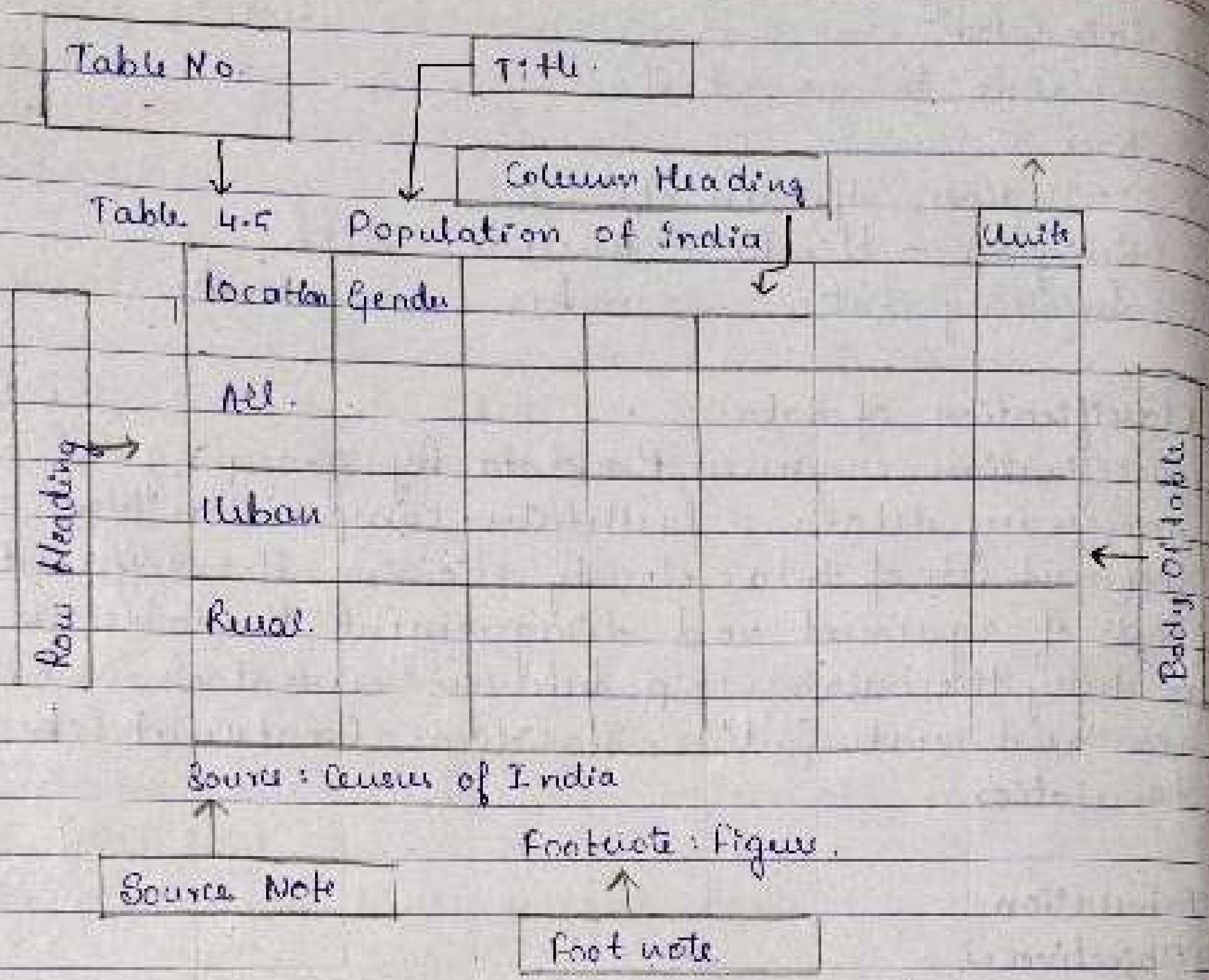
Ex: Hair Color
- blonde, brown red etc
Race
- Indian, African etc
Religion - Hindu, Muslim, Sikh
Smoking status - smoker, non-smoker

* Classification of data.
classification condenses the data by grouping out unnecessary details It facilitates comparision b/w different sets of data clearly showing the different points of agreement and disagreement It enables us to study the relationship between several characteristics and make further statistical treatment like tabulation

* Tabulation
→ Objectives :-
i) To carry out investigation
ii) To do comparision
iii) To simplify data.
iv) To locate omission and errors in data
v) To use space economically
vi) To use it for future reference

* Main parts of a statistical table.
i) Table Number
ii) Title
iii) Column Headings
iv) Stubs / Row Headings
v) Body of the table
vi) Unit of measurement
vii) Source note
viii) Footnote

Structure of the table



Types Of Tables

- Simple table
- Two way table
- Three way table
- Higher order table

1. Draw up a blank table to show the number of employees in a large commercial firm, classified according (i) Sex: Male and female; (ii) Three age-groups: below 30, 30 and above but below 45, 45 and above (iii) four income-groups: below Rs. 400; Rs. 400 750

Rs. 750-1,000, above Rs. 1,000

→ Table No - 01

No. of Employees in firm

Age Group	Income Group												Total
	<400			400 - 750			750-1000			>1000			
	M	F	T	M	F	T	M	F	T	M	F	T	
<30													
30- 45													
≥45													
Grand Total													

Foot note: M - male
F - female
T - total

Source: From ppt

* Draft a blank table to show the population of a town according to (i) Sex: Men and Women (ii) Religion HMC (iii) Wages - Below ₹5000, ₹5000-10,000, ₹10,000 & above.

→ Table No - 02

Population of a town

Wages	Religion																											Total
	H			Mu			C			H			Mu			C			H			Mu			C			
	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T	M	F	T	
Below ₹5000																												
₹5000-₹10,000																												
₹10,000 & above																												

Footnote: H - Hindu
Mu - Muslim
C - Christian
M - Male
F - Female
T - Total

Source:

3 In a sample study regarding smoking habit in a town the following data were obtained
men population - 58%
Smoker - 22%
Men smokers - 18%
Tabulate the above.

→ Table No - 03

Smoking Habit Report.

Pop / Habit	Men	Women	Total
Smoker	18%	4%	22%
Non smoker	40%	38%	78%
Total	58%	42%	100%

Source: Lesson Plan Note.

4. LP → Accident

1960 - 3500 { M 245 { C, NC 49 ; B

1961 - 3200 { M 346 { NC 77, C ; B

1962 - 2800 { M 428 { NC 108, C ; B

1960 —— 3500

- $\xrightarrow{M}$ 245
  - $\xrightarrow{C}$ 196
  - $\xrightarrow{NC}$ 49
- $\xrightarrow{B}$ 3255
  - $\xrightarrow{C}$ 2,867
  - $\xrightarrow{NC}$ 388

1961 —— 3200

- $\xrightarrow{M}$ 346
  - $\xrightarrow{C}$ 269
  - $\xrightarrow{NC}$ 77
- $\xrightarrow{B}$ 2854
  - $\xrightarrow{C}$ 2587
  - $\xrightarrow{NC}$ 267

1962 —— 2800

- $\xrightarrow{M}$ 428
  - $\xrightarrow{C}$ 320
  - $\xrightarrow{NC}$ 108
- $\xrightarrow{B}$ 2372
  - $\xrightarrow{C}$ 2[illegible]52.
  - $\xrightarrow{NC}$ 220

→ Table No - 04.

Number of Accidents in Southern Railway from 1960 - 1962.

Year	Accidents				Total.
	Metre Gauge		Broad Gauge		
	Compensated	Non-Compensated	Compensated	Non-Compensated	
1960	196	49	2867	388	3500
1961	269	77	2587	267	3200
1962	320	108	2152	220	2800

Frequency Distribution

* When dataset contains more than 50 items group the data and use statistical measures on data
* The steps in preparing grouped frequency distribution are:- i) Determining the class intervals
  ii) No. of intervals $n = 1 + 3.322 \log \frac{N}{10}$.
  N - no. of observations in dataset
  is called Sturge's formula
  iii) Width of the interval

$$w = \frac{UL - LL}{n}$$

where UL - upper limit
LL - lower limit

2) iv) Recording the data using tally marks
3) v) Finding frequency of each class by counting the tally marks

* The numbers in the frequency column show how many items fall into each class and they are called the frequency of those classes
* The width of the class is called the class interval
* The midpoint of class is called the classmark
* A set of rawdata summarized by distributing it into a number of classes along with their frequencies is known as a frequency distribution

Percentage frequency distribution:-

Percentage frequency of class interval =

$$\boxed{\underbrace{\frac{\text{Class frequency}}{\text{Total freq}}}_{\text{Relative frequency}}} \times 100$$

Cummulative frequency distribution:-

Cummulative frequency of a class interval can be obtained by adding the frequency of that class interval to the sum of the frequencies of the preceding class interval.

Exclusive class interval (Upper limit is not included & lower limit is included in the interval).

C.I	freq	c.f
10-20	2	2
20-30	6	8
30-40	8	16
40-50	4	20

Inclusive class interval

C.I	freq
10-19	
20-29	
30-39	
40-49	

Conversion to Exclusive
(diff of UL and LL of next CI %2)
Add diff to UL
Sub diff from LL.

C.I
9.5-19.5
19.5-29.5
29.5-39.5
39.5-49.5

* Numerical Method for summarising Quantitative Data

1) Measures of Central Tendency (mean, median, mode).
2) Measures of variation (Measures of dispersion) (SD (mean), Quartile deviation).

A frequency distribution shows clustering of the data around some central value differed methods give ______ diff averages which are known as the measure of central tendency
The commonly used measures of central values are mean, median, mode.

Mean (Arithmatic average): Mean of a set of numbers is computed by adding all the values in the data set and divide by the no. of observations

$$\text{Mean} = \bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\Sigma x_i}{n}$$

n - no of observation

In a frequency distribution

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{n} = \frac{\Sigma x_i f_i}{n}$$

C.I	f	x_i
10 - 20	2	15
20 - 30	6	25
30 - 40	8	35
	$\Sigma f_i = 16 = n$	

Use the mean to describe middle of set of data that does not have an outlier.

Advantages:- * Most popular measure in fields such as business, engineering and computer science
* It is unique
* Useful when comparing sets of data

Disadvantages:-
* Affected by extreme values (outliers)
Outliers are extreme or typical data values that are notably different from the rest of data

* Median the median is the middle value in distribution when the values are arranged in ascending or descending order.

Use the median to describe the middle value that does have an outlier

Advantages:
* Extreme values do not affect the median as strongly as do the mean
* It is unique
* Useful when comparing sets of data

Disadvantages:
* Not popular as mean

Medians for grouped data.

$$\text{Median} = l + \frac{(N/2 - c)}{f} \times h$$

l- lower limit of the median class.
N- total no. of observations
c- cummulative freq upto the class frequency the median class.
f- freq of the median class
h- width of the median class.

Mode: Mode is most commonly occuring value in distribution. Use the mode when the data is non-numeric or when asked to choose the popular item.

Advantages:

* Extreme values do not affect the mode.

Disadvantage:

* Not popular as mean and median
* Not necessarily unique
* When no values repeat in the dataset the mode is every value and is useless
* When there is more than one mode it is difficult to interpret and compare.

For grouped data,

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

where l - lower limit or the model class.
f_1 - freq of the class in which mode lies
f_0 - freq of the class preceding model class
f_2 - freq of the class succeding model class.
h - width of the model class

Ex: 1) weights (in kg) students of class:

42, 72, 40, 60, 82
115, , 41, 61, 75, 83
63, 53, 110, 76, 84
50, 67, 65, 78, 77
56, 95, 68, 69, 104
80, 79, 49, 54, 73
59, 81, 110.

Sturges formula

$n = 1 + 3.322 \log_{10} N$
$= 1 + 3.322 \log_{10} 33$
$= 6.04 = \underline{6}$

$h = \frac{UL - LL}{n} = \frac{115 - 40}{6} = 12.5$
$\simeq 13.$

x_i	C.I	freq	Tally	c.f	$f_i x_i$
46.5	40 - 53	4	IIII	4	186
59.5	53 - 66	8	~~IIII~~ III	12	476
72.5	66 - 79	9	~~IIII~~ IIII	21	652.5
85.5	79 - 92	7	~~IIII~~ II	28	598.5
98.5	92 - 105	2	II	30	197
111.5	105 - 118	3	III	33	334.5

$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{244.5}{33} = 74.075$

$N = 33$

$\frac{N}{2} = \underline{16.5}$ The 17^{th} observation lies in the interval 66 - 79

$\text{Median} = L + \left(\frac{\frac{1}{2}N - c}{f}\right) \times h = 72.5$

$l = 66$
$N = 33$
$C = 12$
$f = 9$
$h = 13$

The modal class is 66 - 79 which has highest frequency.

$$\text{Mode} = L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times h$$

$$= 66 + \frac{9-8}{(1)+(9-7)} \times 13$$

$$= 66 + \frac{1}{3} \times 13 = 70.33$$

(P 9:-

$$\text{Mode} = l + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times h.$$

$$= 66 + \frac{9-8}{(1) + (9-7)} \times 13$$

$$= 66 + \frac{1}{3} \times 13 = 70.33.$$

(P9:- The following are the grades of 50 students in statistics class.

75	89	66	52	90	68	83	94	77	60
38	47	87	65	97	49	65	70	73	81
85	77	83	56	63	79	69	82	84	70
62	75	29	88	74	37	81	76	74	63
69	73	91	87	76	58	63	60	71	82

$\rightarrow n = 1 + 3.322 \log_{10} 50 = 6.64 = 7.$

$$h = \frac{ul - ll}{n} = \frac{97 - 29}{7} = 9.7 = 10.$$

C.I.	Tally	f	cf	x_i	$x_i f_i$
29-39	III	3	3	34	102
39-49	I	1	4	44	44
49-59	IIII	4	8	54	216
59-69	~~IIII~~ ~~IIII~~	10	18	64	640
69-79	~~IIII~~ ~~IIII~~ ~~IIII~~	15	33	74	1110
79-89	~~IIII~~ ~~IIII~~ II	12	45	84	1008
89-99	~~IIII~~	5	50	94	470

$$\text{mean} = \frac{\Sigma f_i x_i}{N} = \frac{3590}{50} = 71.8$$

$$\text{median} = 69 + \frac{\frac{1}{2} \times 50 - 18}{15} \times 10$$

$$= 69 + \frac{15}{15} \times 10$$

$$= 73.666$$

$$\text{Mode} = 69 + \frac{(15-10)}{(15-10) + (15-12)} \times 10$$

$$= 69 + \frac{5}{8} \times 10$$

$$= 75.25$$

* Cummulative frequency distribution.

We calculate the cummulative frequencies by adding the relative frequencies as we go down the class and this will generate the ogive line.

(i) Less than ogive: Plot the points with the upper limit of the classes on x-axis and the corresponding less than cumulative frequency on y-axis. Join the points by a free hand and smooth curve to get less than ogive and it is a raising curve.

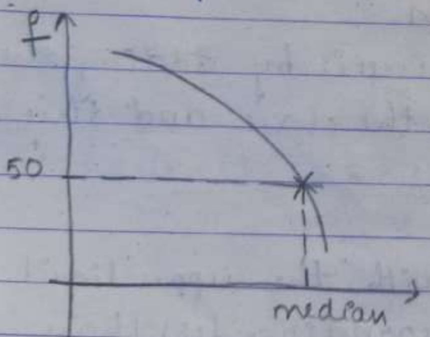
(ii) More than ogive: Plot the points with lower limits of the classes on x-axis and the corresponding

more than the cummulative frequency on y-axis. Join the points by free hand smooth curve to get more than ogive it is falling curve
To

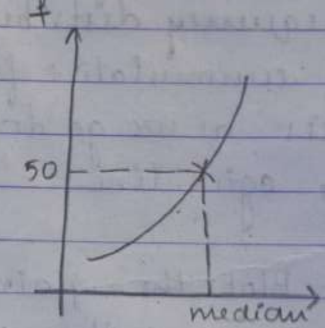
NOTE: The ogives give a ready method of making on the curve the values of the median and quartiles

The two ogives less than and more than cut each other at the median.

How to construct ogive?
→ The cummulative frequency of each class is plotted against the upper limit of the class interval for less than ogive.
→ lower limit of the class interval is used for more than ogive

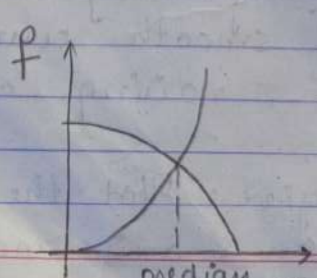


CI (more than ogive)



CI (less than ogive)

FALLING CURVE

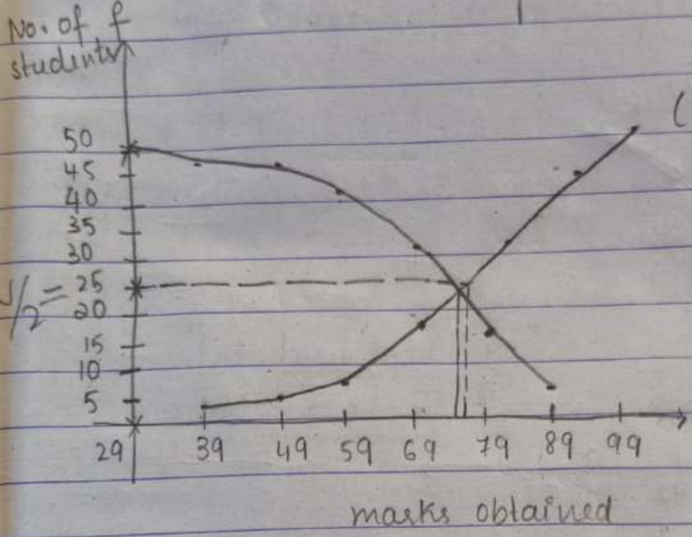


Less than ogive

Marks less than	f(cf)
39	3
49	4
59	8
69	18
79	32
89	45
99	50

More than ogive

marks more than or equal	f
29	50
39	50-3=47
49	47-1=46
59	46-4=42
69	42-10=32
79	32-15=17
89	17-12=5



(U.L) Rising curve less than ogive.

C.I

from less than ogive median is 74.

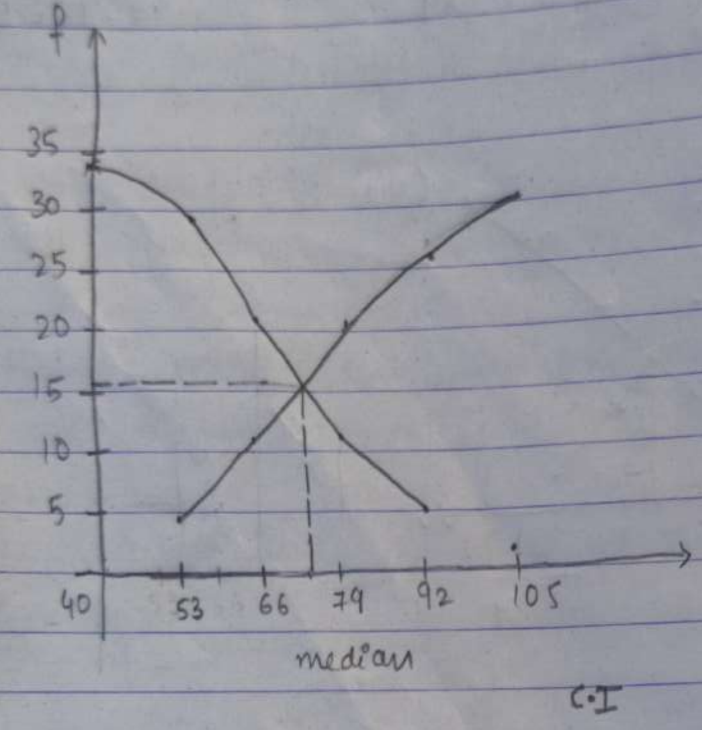
More than ogive (cummulative frequency curve).

More than ogive

Less than ogive

Marks less than	c.f.
53	4
66	12
79	21
92	28
105	30
118	33

Marks more than	f
40	33
53	33 - 4 = 29
66	29 - 8 = 21
79	21 - 9 = 12
92	12 - 7 = 5
105	5 - 2 = 3



* Measures of dispersion.

Although measures of central tendency do exhibit one of the important characteristics of distribution yet they fail to give any idea as to how the individual value differ from central value i.e whether they are closely packed around the central value or widely scattered away from it

Two distribution may have the same mean and total frequency yet they may differ in the extent to which the individual values may be spread above the average. The magnitude of such variation is called dispersion

The following are the measures of dispersion:

1) Range: Defined as a single number representing the spread of the data

Range = Upper value - Lower value

2) Standard Deviation (S.D). It is defined as a number representing how far from the average each score is. It is the important and powerful measure of dispersion. and is denoted by σ

For raw data

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

For grouped data.

$$\sigma = \sqrt{\frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i = N}}$$

The square of the S.D. is known as variance

$$\text{Variance} = V = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

Co-efficient of variation: It is the percentage variation in the mean, standard deviation being considered as total variation in the mean

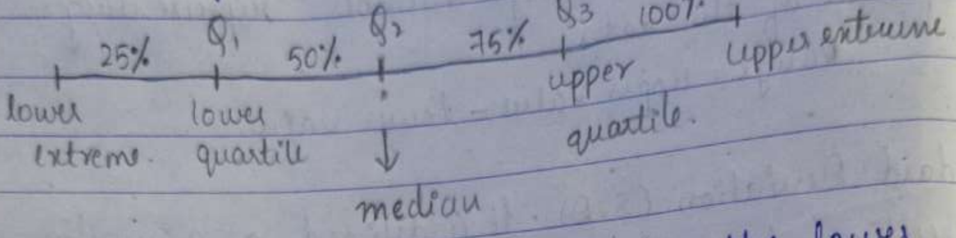
$$C.V = \frac{\sigma}{} \times 100.$$

$\bar{x}$ = mean.

σ = S.D.

3) Quartile Deviation:

Quartiles are those value which divide the frequency into four equal parts when the values are arranged in the ascending order of magnitude.



Lower Quartile Q_1 is midway b/w the lower extreme and median.

Upper Quartile Q_3 is midway b/w median and upper extreme.

For the grouped data

$$Q_1 = L + \frac{(N/4 - c)}{f} \times h.$$

$$Q_3 = L + \frac{(3N/4 - c)}{f} \times h.$$

c - cummulative frequency of preceding class

f -

Quartile Deviation is one half of the inter-quartile range i.e. Quartile deviation (Q.D)

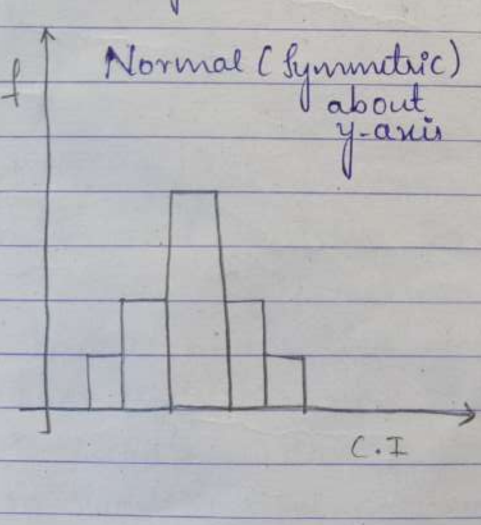
$$Q.D. = \frac{1}{2}(Q_3 - Q_1)$$

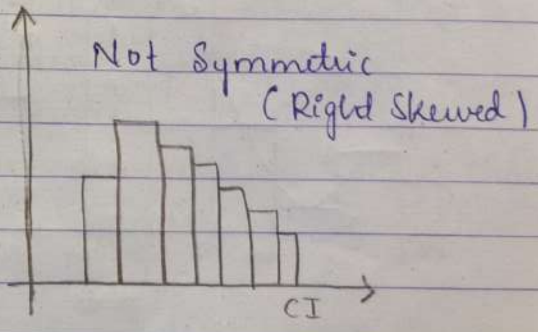
Histogram: A convienient way of representing a sample frequency distribution is by means of graphs. It gives the general run of the observation. A histogram is drawn by erecting rectangles over the class interval such that areas of the rectangles are proportional to the class frequencies. If the CI are of equal size the height of the rectangles will be proportional to the class frequencies.

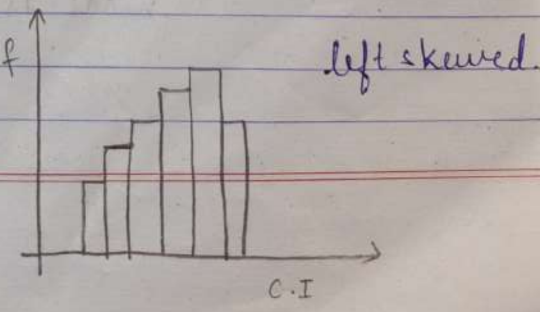
Drawing Histogram:

Mark off along the x-axis all the CI on the suitable scale.

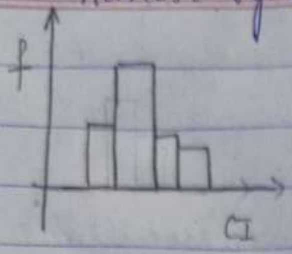
- Mark frequencies along y-axis on suitable scale
- We can have different scales for the two axes.
- Construct a rectangles with CI as bases and heights proportional to frequencies





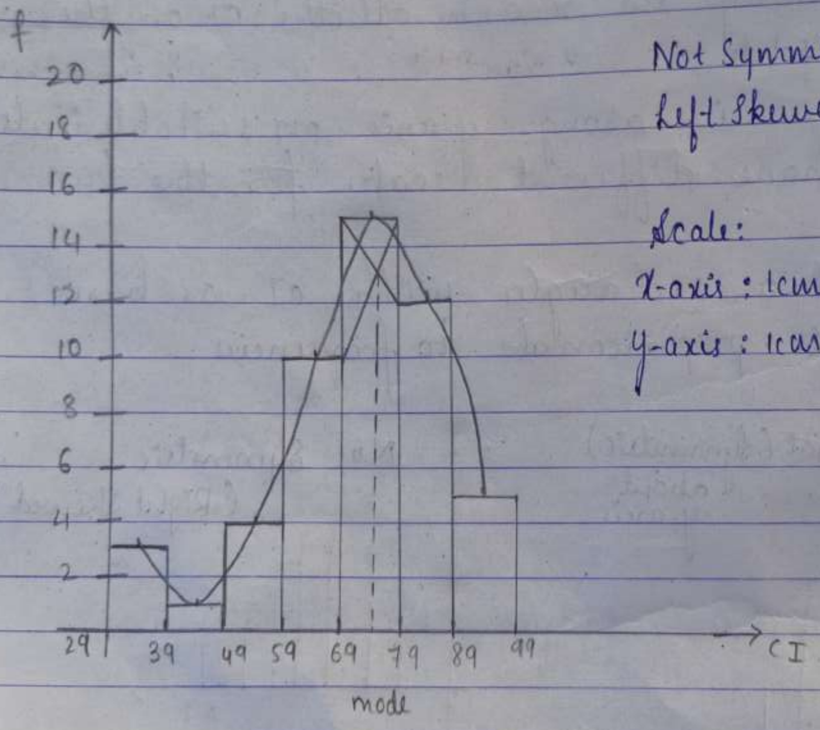


Almost Symmetric



LP9:- Compute mode analytically and graphically

C.I	f
29 - 39	3
39 - 49	1
49 - 59	4
59 - 69	10
69 - 79	15
79 - 89	12
89 - 99	5



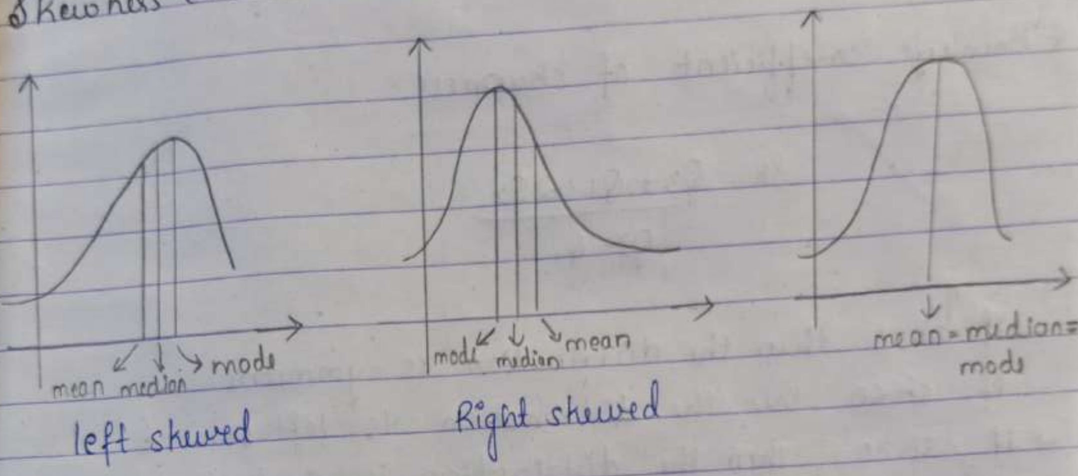
Not Symmetric
Left Skewed

Scale:
X-axis : 1cm = 10 units
Y-axis : 1cm = 2 units

mode = 75

* Frequency Curve: For a grouped freq dist'n with equal class intervals a freq curve is obtained by joining the middle points of upper side (tops) of the adjacent rectangles of the histogram by means of smooth curve. (If we join the middle points by a straight line gives frequency polygon)

Skewness (*)



Skewness measures the degree of symmetry. If the frequency curve has a longer tail to the right i.e. the mean is to the right of the mode. Then the distribution is said to have positive skewness (right skewed)

If the frequency curve is more elongated to the left then it is said to have negative skewness (left skewed)

skewness means lack of symmetry. A distribution is said to be skewed if

i> mean ≠ median ≠ mode
ii> Quartiles are not equidistant from median.
iii> The curve drawn with help of the given data is not

symmetric but streched more to one side than to the other.

* Karl Pearson's co-efficient of skewness

- Denoted by SK

$$Sk = \frac{\text{mean} - \text{mode}}{\text{S.D}}$$

* Bowley's co-efficient of skewness.

$$Sk = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

→ if $sk = 0$, then the distribution is symmetric
→ if $sk < 0$, then the distribution is left skewed
→ if $sk > 0$, then the distribution is right skewed.

Examples:-

[P 9] mean = 7[illegible].[illegible]8 = $\bar{x}$
mode = 75.25

CI	f	xi	xifi	$fi(xi-\bar{x})^2$
29-39	3	34	102	4286.52
39-49	1	44	44	772.84
49-59	4	54	216	1267.36
59-69	10	64	640	608.4
69-79	15	74	1110	72.6
79-89	12	84	1008	1786.08
89-99	5	94	470	2464.2
	$\Sigma fi = 50$		3590	11258

$$S.D = \sqrt{\frac{\Sigma f_i (x_i - \bar{x})^2}{N/\Sigma f_i}}$$

$$= \sqrt{\frac{11258}{50}} = 15.00$$

By using Karl Pearson's co-efficient of skewness

$$Sk = \frac{\text{mean} - \text{median}}{SD.}$$

$$Sk = \frac{71.8 - 75.25}{15}$$

$$= -0.23$$

Negatively skewed (left skewed)

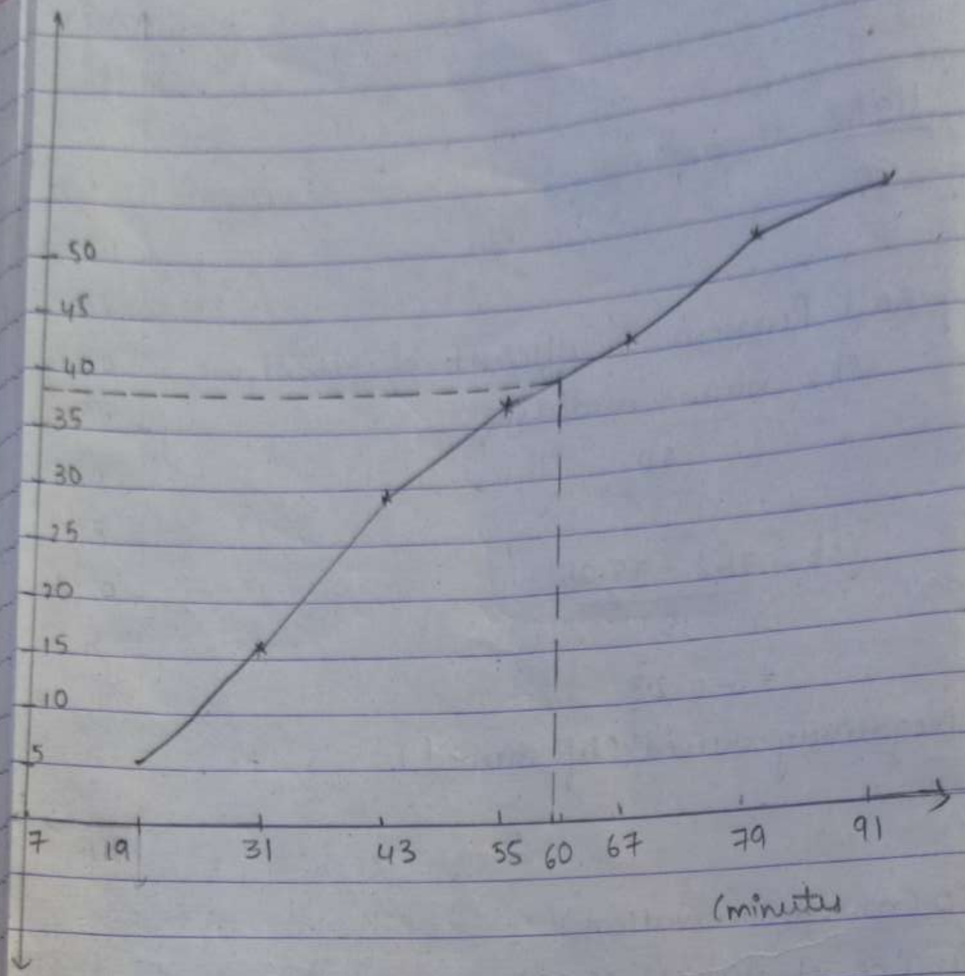
Q11) N=50 (no. of observations)

n= 7 (no. of classes given)

$$h = \frac{UL - LL}{n} = \frac{88-7}{7} = \frac{81}{7} = 11.57 \simeq 12$$

C.I	Tally	f	cf
7 - 19	~~IIII~~ I	6	6
19 - 31	~~IIII~~ ~~IIII~~	10	16
31 - 43	~~IIII~~ ~~IIII~~ III	13	29
43 - 55	~~IIII~~ III	8	37
55 - 67	~~IIII~~	5	42
67 - 79	~~IIII~~ I	6	48
79 - 91	II	2.	50

Ogive (less than ogive)



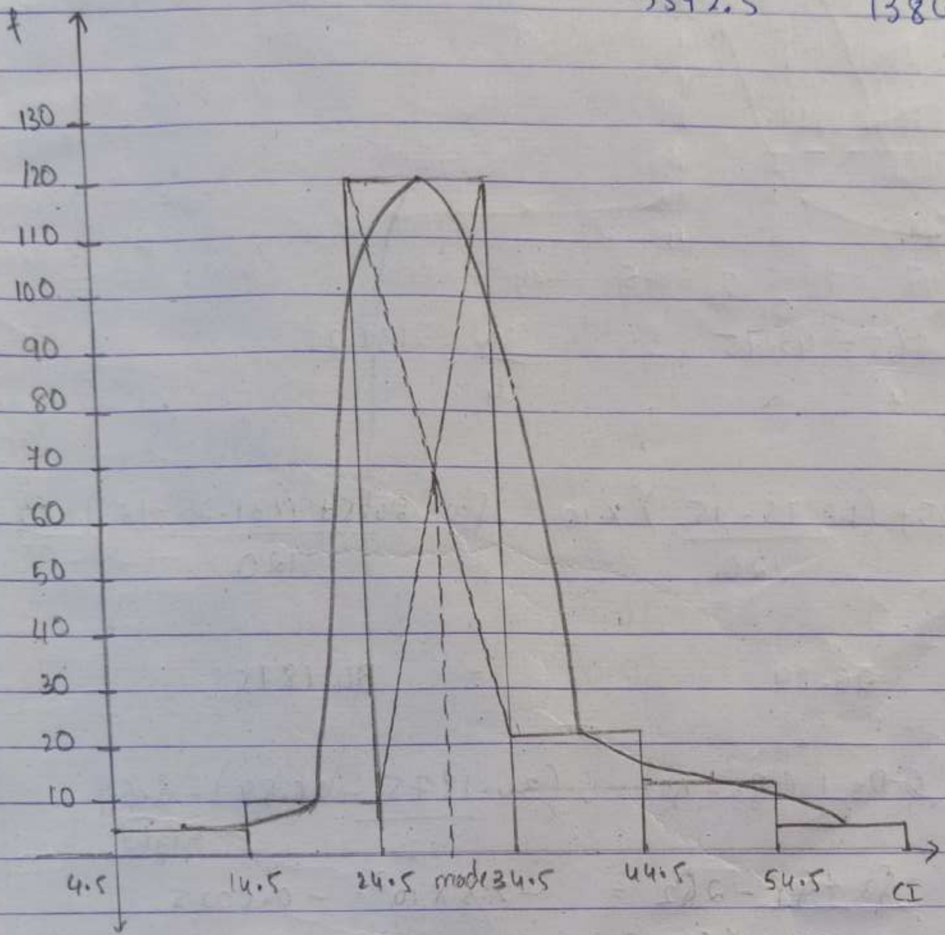
From ogive, we can see that about 37 students subscribers spent 60 minutes or less online during their last session

The greatest increase in usage occurs between 31-43 minutes because the line segment is steepest b/w these two boundaries

Q 10] Making the CI exclusive.

C.I (Age)	4.5 - 14.5	14.5 - 24.5	24.5 - 34.5	34.5 - 44.5	44.5 - 54.5	54.5 - 64.5
No. of classes/freq	5	10	120	22	13	5

C.I	f	cf	xi	fixi	$f_i(x_i-\bar{x})^2$
4.5 - 14.5	5	5	9.5	47.5	2520.01
14.5 - 24.5	10	15	19.5	195	1550.025
24.5 - 34.5	120	135	29.5	3540	720.3
34.5 - 44.5	22	157	39.5	869	1254.055
44.5 - 54.5	13	170	49.5	643.5	4004.0325
54.5 - 64.5	5	175	59.5	297.5	3795.0125
				5592.5	13843.435



$\sigma = \sqrt{\frac{\sum f_i(x_i-\bar{x})}{\sum f_i}} = 8.89$

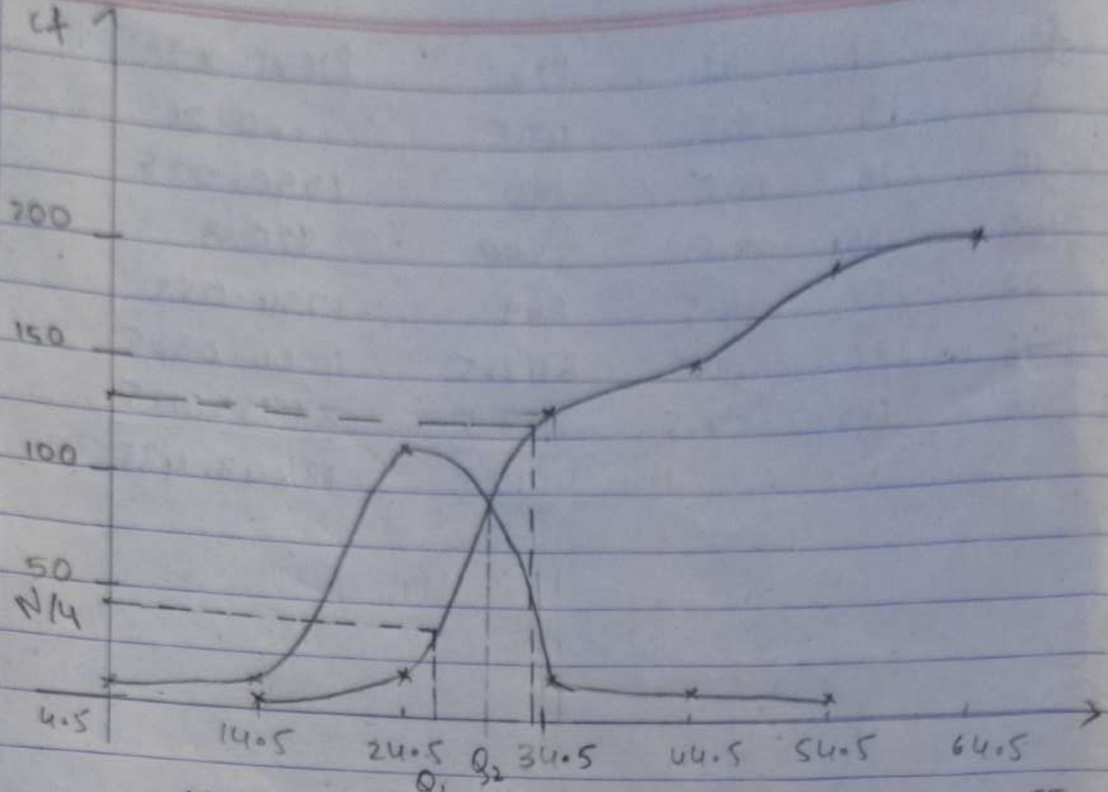
$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = 31.9571$

variance = 79.0321

median = $24.5 + \frac{87.5 - 15}{120} \times 10 = 30.5416$

$= 29.7884$

mode = $24.5 + \frac{(120 - 10)}{110 + (120 - 22)} \times 10$



$\frac{N}{4} = \frac{175}{4} = 43.75$ $\qquad$ $\frac{3N}{4} = 131.25$

$Q_1 = 24.5 + \left(\frac{43.75 - 15}{120}\right) \times 10$ $\qquad$ $Q_3 = 24.5 + \left(\frac{131.25 - 15}{120}\right) \times 10$

$= 26.89$ $\qquad$ $= 34.1875$

Karl Pearson $Q.D = \frac{1}{2}(Q_3 - Q_1) = \frac{(34.1875 - 26.89)}{2} = 3.64$

right

Bowley : - $\frac{Q_3 + Q_1 - 2Q_2}{7.2975\,(Q_3 - Q_1)} = -2.5 \times 10^3 = -0.0025$

left

It is right skewed not symmetric

→ By Karl Pearson's co-efficient of skewness.

$SK = \frac{mean - mode}{SD} = 0.24 > 0$ so +ve & right skewed

(P 12) N = 30 days = 30

By Strugi's formula

$n = 1 + 3.322 \log_{10} N$

$= 1 + 3.322 \log_{10} 30$

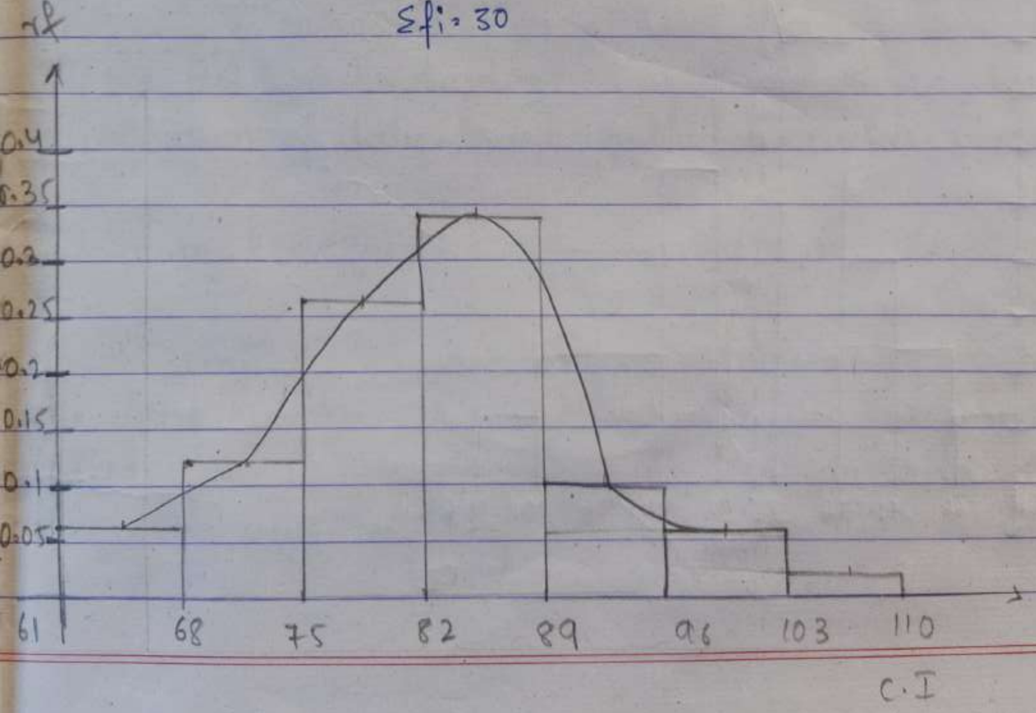
$= 1 + 3.322$

$= 5.906 \simeq 6$

$h = \frac{UL - LL}{n} = \frac{104 - 61}{6} = 7.16 \simeq 7$

Take eight classes

C.I	Tally	f	r.f
61-68	‖	2	0.06
68-75	‖‖	4	0.13
75-82 |卌 |||	8	0.26
82-89	卌 卌	10	0.33
89-96	|||	3	0.1
96-103	‖	2	0.06
103-110			1	0.03
		$\Sigma f_i = 30$	

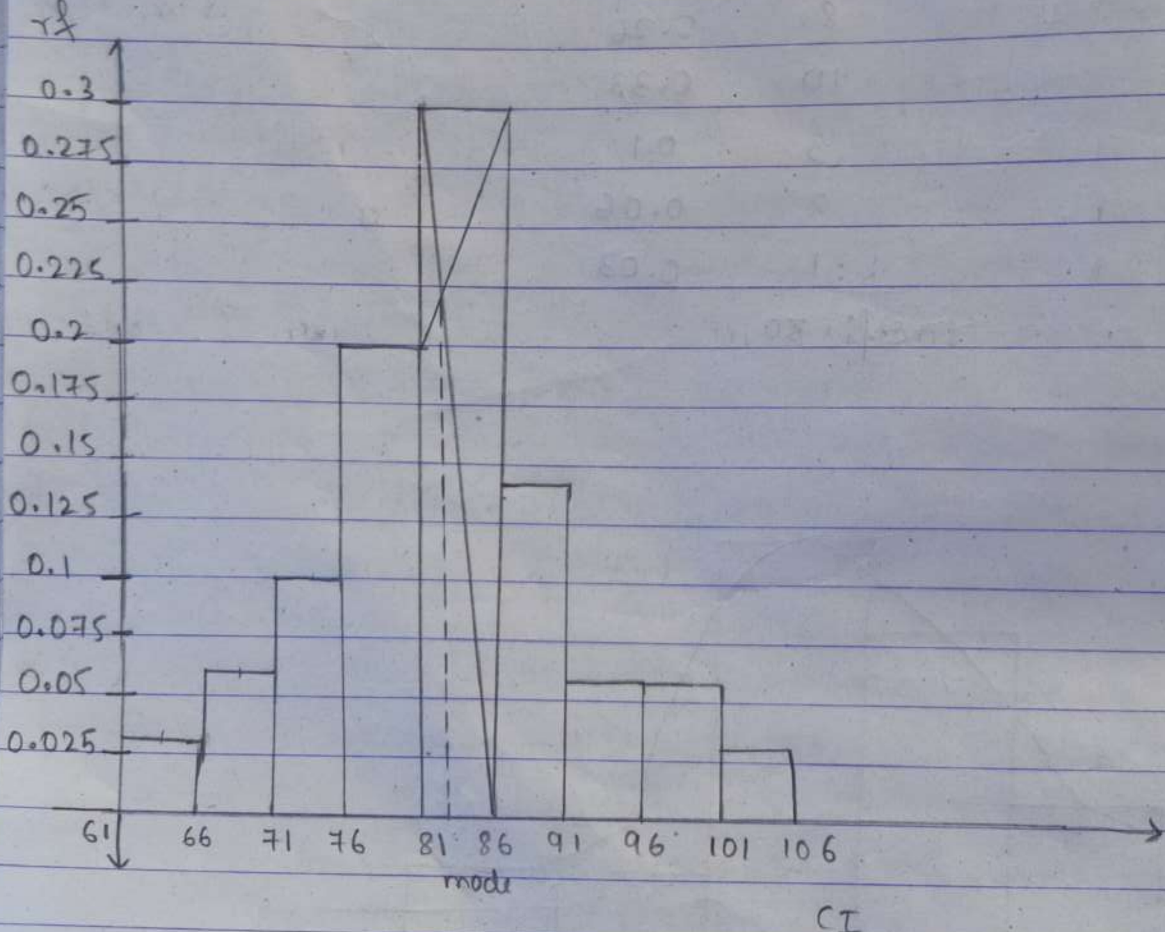


$n = 8$

$h = \frac{UL - LL}{n} = \frac{104 - 61}{8} = 5.375 \approx 5$

(100's dollar)

C.I	Tally	f	rf	cf
61-66	|	1	0.03	1
66-71	||	2	0.06	3
71-76	|||	3	0.1	6
76-81	~~||||~~ |	6	0.2	12
81-86	~~||||~~ ||||	9	0.3	21
86-91	||||	4	0.133	25
91-96	||	2	0.066	27
96-101	||	2	0.066	29
101-106	|	1	0.033	30.



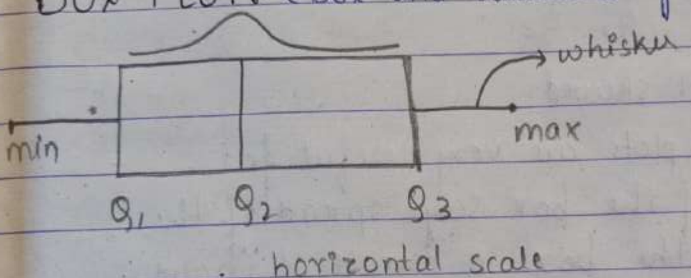
12b) \$9000 is in class interval 86-91. So after that there will be run out of cash after that. So add all the r f after that.

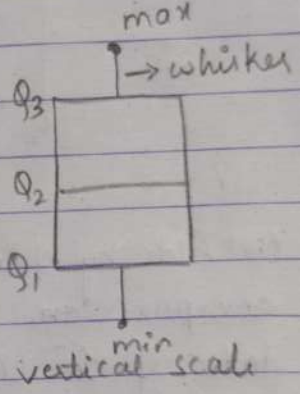
16.7% will be run out of cash if we put \$9000 in the atm each day ∴ the sum of the relative frequency for the last three classes 0.167 = 16.7%.

12c) If we add last two relative frequency we get 10%.

\$9600 if we put to run out of cash for 10% ∴ the sum of relative frequency for the last two classes is 0.099 ≃ 0.1

BOX PLOT:- (Box and Whisker Diagram)





The box plot is a standardized way of displaying the distribution of data based on the 5 number summary

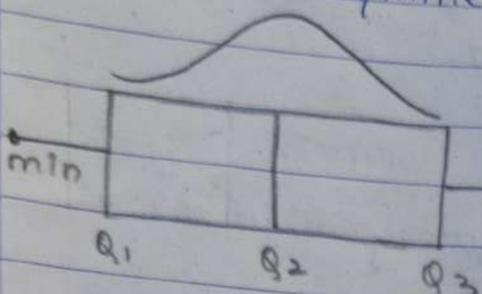
Minimum, First Quartile, Median, Third Quartile, Maximum

In the simplest Box plot the central rectangle spans the first quartile to the third quartile. A segment inside the rectangle shows the median and whiskers left and right of the box.

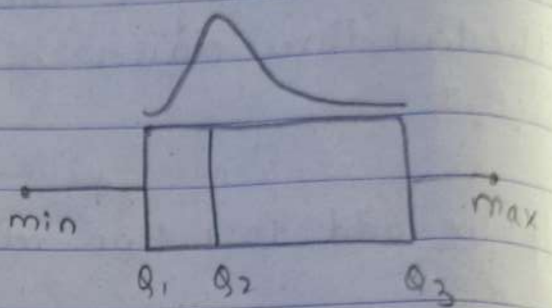
Histogram - grouped data
box plot - raw data (better)

Merits:-

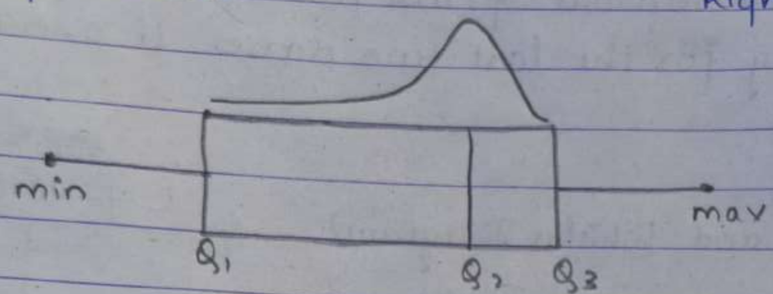
(i) Much more compact than histogram
(ii) Quick visual picture
(iii) Gives rough idea on how data is distributed, position of the median line indicates symmetric or non-symmetric



Symmetric



Right-skewed



Left-skewed.

(iv) Side by side box plots are very useful for comparision. Size of the box says spread of the data if the size of the box is small it shows data's are having same opinion.
If size is big shows data's are having different opinion

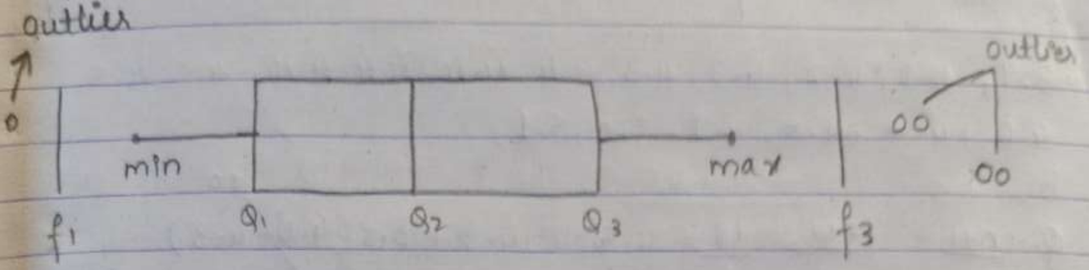
Construction:

(i) Arrange data in ascending order
(ii) find the sample median Q_1, Q_3
(iii) Find two points f_1 and f_3 called inner fences

$f_1 = Q_1 - 1.5\,(IQR)$ IQR - Inter Quartile Range

$f_1 = Q_1 - 1.5\,(Q_3 - Q_1)$

$f_3 = Q_3 + 1.5(Q_3 - Q_1)$



These points will be used to identify the outliers.

$Q_1 = \frac{n+1}{4}$ → position of first quartile.

$Q_2 = \frac{n+1}{2}$ → position of the median.

$Q_3 = \frac{3}{4}(n+1)$ → position of the third quartile.

NOTE:-

$Q_1 = 7.5$th position
$= 7$th position value $+ 0.5[8\text{th} - 7\text{th}]$.

$Q = 6.25$th position
$= 6\text{th} + 0.25(7\text{th} - 6\text{th})$.

Example 1:-

Draw a box plot for the following data set.

4.3, 5.1, 3.9
4.5, 4.4, 4.9
5.0, 4.7, 4.1
4.6, 4.4, 4.3
4.8, 4.4, 4.2
4.5, 4.4.

(i) Step1: Ascending order of the given sample.

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5,
4.6, 4.7, 4.8, 4.9, 5.0, 5.1.

$Q_1 = \frac{n+1}{4} = \frac{17+1}{4} = \frac{18}{4} = \underline{\underline{4.5}} = 4.3 + 0.5(\cancel{4.3 - 4.3})^{\nearrow 0} = 4.$

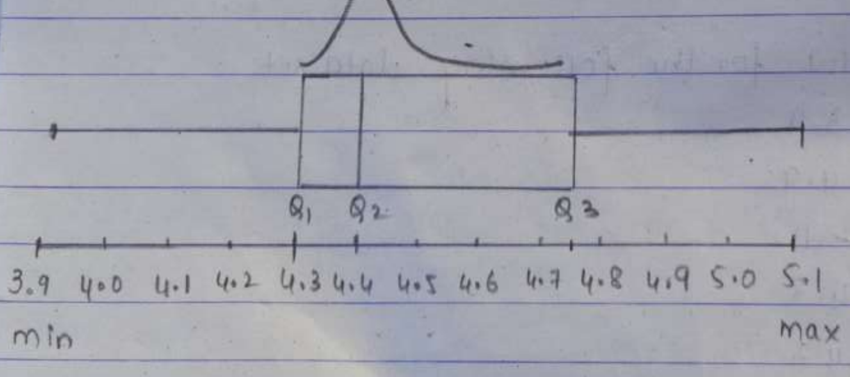
$Q_2 = \frac{n+1}{2} = \frac{17+1}{2} = \frac{18}{2} = \underline{\underline{9}} = \underline{\underline{4.4}}$

$Q_3 = \frac{3}{4}(n+1) = \frac{3}{4} \times 18 = 13.5 = 4.7 + 0.5(4.8 - 4.7)$
$= 4.75.$

$f_1 = Q_1 - 1.5(Q_3 - Q_1)$
$= 3.625$

$f_3 = Q_3 + 1.5(Q_3 - Q_1)$
$= 5.425$

There are no values less than 3.625 and greater than 5.425. There are no left or right outlier ∴ no values are less than 3.625 or greater than 5.425



Q) Let 'x' denote the difference in temperature b/w the surface of water and the water depth of 1 km. Measurements are taken at 15 randomly selected sites in the Gulf of Mexico. These data result in the following temperature. Draw box plot and discuss symmetry and outliers.

22.5, 23.8, 23.2, 22.8

10.1, 23.5, 24.0, 23.2

24.2, 24.3, 23.3, 23.4

23.0, 23.5, 22.8

⇒ Step 1: Arrange in ascending order

10.1, 22.5, 22.8, 22.8, 23.0, 23.2, 23.2, 23.3, 23.4

23.5, 23.5, 23.8, 24.0, 24.2, 24.3

$Q_1 = \frac{n+1}{4} = \frac{16}{4} = 4$ th position = 22.8

$Q_2 = \frac{n+1}{2} = \frac{16}{2} = 8$ th position = 23.3

$Q_3 = \frac{3}{4}(n+1) = \frac{\cancel{16}^{4} \times 3}{\cancel{4}} = 12$ th position = 23.8

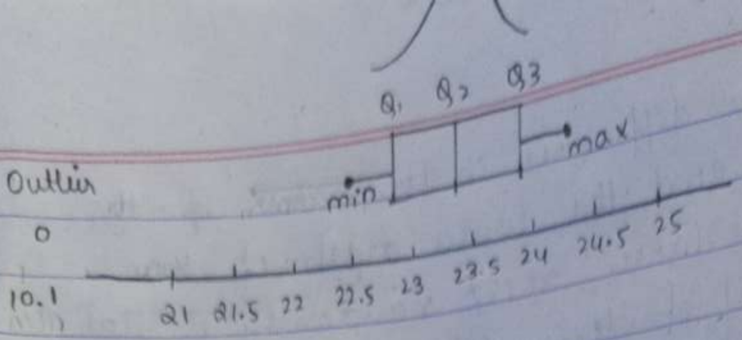
$f_1 = Q_1 - 1.5(Q_3 - Q_1)$

$= 21.3$

$f_2 = Q_3 + 1.5(Q_3 - Q_1)$

$= 25.3$

There are values less than 21.3 i.e 10.1 is the outlier.



Bowley's : $\frac{Q_3 + Q_1 - 2Q_2}{(Q_3 - Q_1)} = \frac{0}{1} = 0$

Distribution is symmetric.

Quantile - Quantile Plot (Q-Q plot)

The Quantile-Quantile Plot is a graphical technique for determing if two datasets come from population with common distribution.

A Quantile - Quantile plot is a plot of the quantile of the first data set against the quantiles of the second data set

Q-Q plot allow us to compare the quantiles of the two sets of numbers. This kind of comparision is much more detailed than the simple comparision of means and medians.

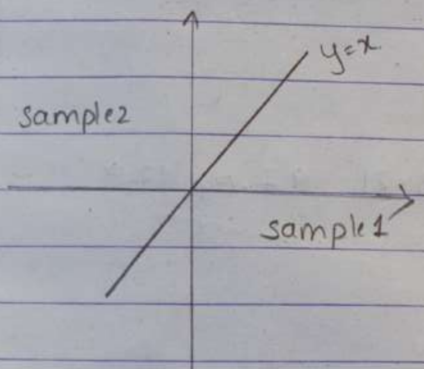
Q-Q plot is used to check
(i) whether the two data sets come from population with common distribution
(ii) whether the two data sets have common location and scale
(iii) whether two datasets have similar distn shapes.

(iv) where the two data sets have similar tail behaviour.

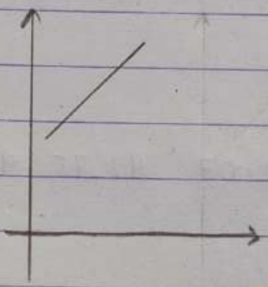
1) If the two dist'n being compared are identical then Q-Q plot follows the line $y=x$.

2) If the two dist'n agree after linearly transformating the values in one of the dist'n then Q-Q plot follows some line but not necessarily the line $y=x$.
(i.e two samples from similar dist'n which differ only in location).

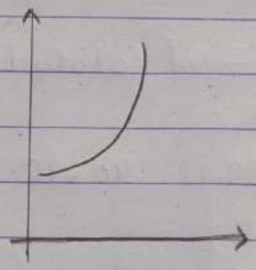
3). If Q-Q plots are S-shaped indicating that one of the dist'n is more skewed than the other or one of the dist'n has heavier tails than the other



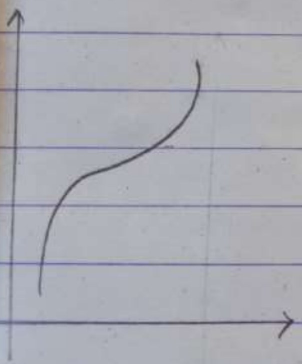
(i) Come from same dist'n identical



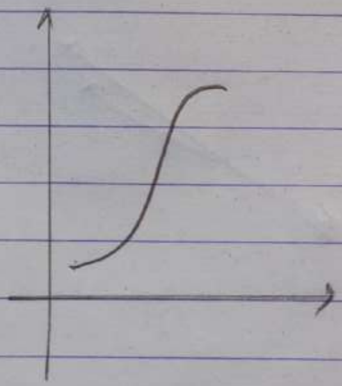
(ii) change in Location



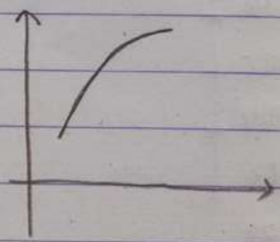
(iii) Right skewed



(iv) Heavy tail



(v) light tail.



(iv) left skewed.

* Construct Q-Q plots Estimate Quantiles from data set 1 and take those values along y-axis

(ii) Estimate Quantiles from dataset 2 and take those value on x-axis

Both axis are in units of their dataset.

Q-Q plot is used to check whether the two data sets come from the same dist'n or not

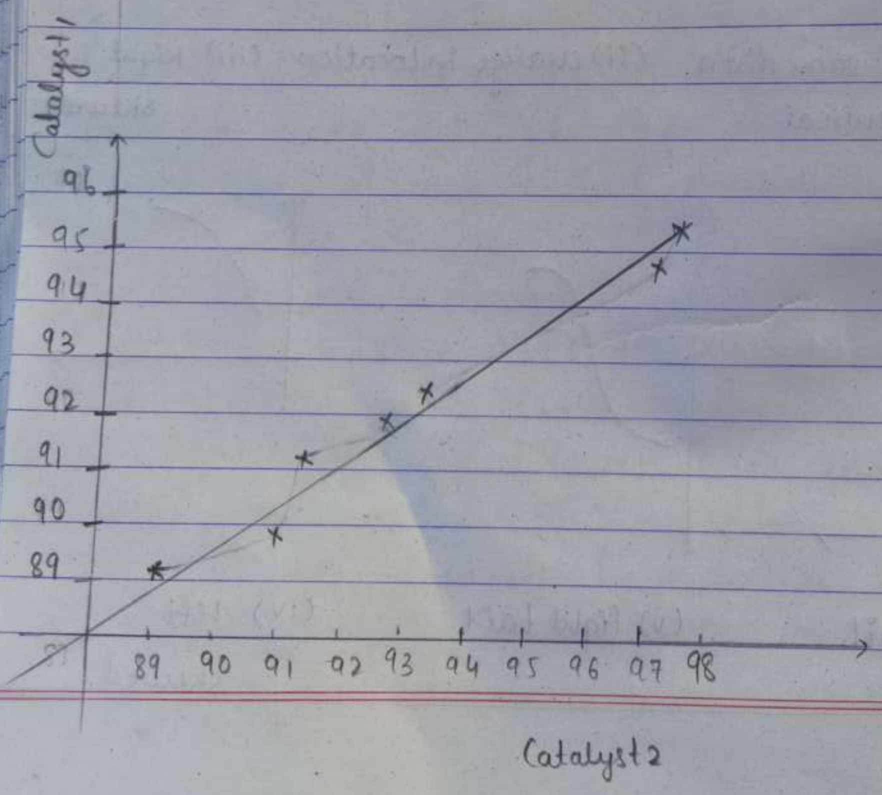
[P15].

Sorted Catalyst 1

89.07 89.21 91.5 91.79 92.18 94.72 95.39

Sorted Catalyst 2

89.18 90.95 91.07 92.75 93.21 97.04 97.19



Normal Quantile Quantile Plot (qq norm)

It is used to check whether the given distribution or given dataset come from normal distribution or not.

Construction of qq norm:

1) First order the data in ascending order

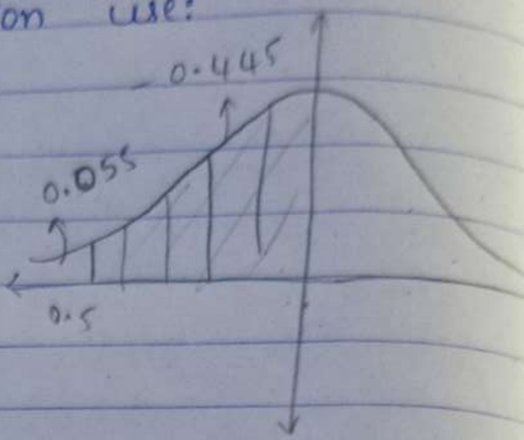
2) Plot these values against appropriate quantiles from the standard normal distribution.

$$Z_i = \frac{i-0.5}{n} \qquad i = \text{Index}$$

3) If points roughly lies on the linearline then given set comes from normal distribution.

To find linear equation use:

$$Z = \frac{1}{\sigma} x - \left(\frac{\bar{x}}{\sigma}\right)$$



$$\text{ll } Z = \frac{x - \mu\,(\text{mean})}{\sigma - \text{S.D}}$$

* Example - 01

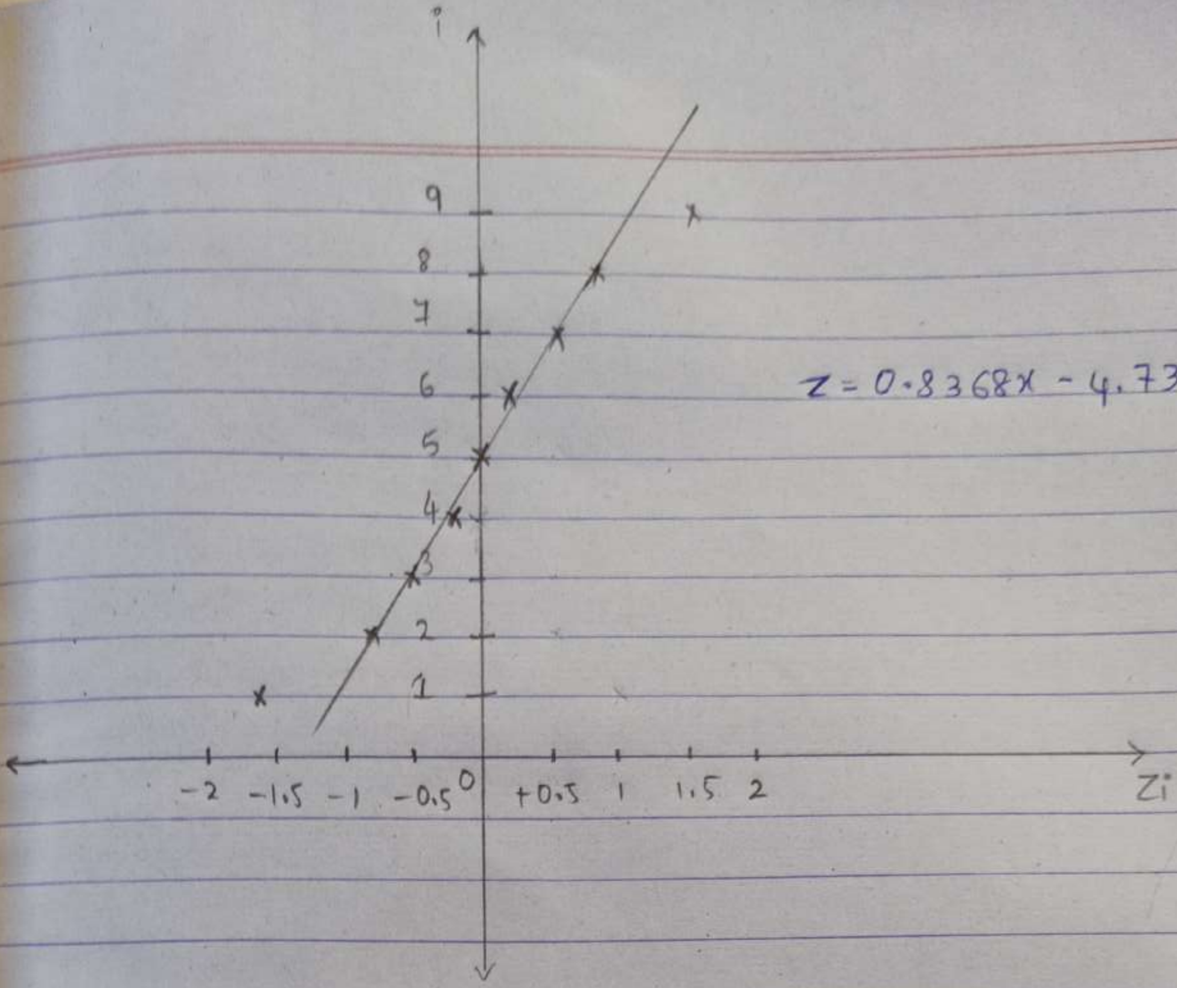
1) Does the following data come from normal dist'n population:

3.89, 4.75, 6.33, 4.75, 7.21, 5.78, 5.80, 5.20, 7.90

i	x_i	$q = i - 0.5 / n = 9$	$A(x_i) = q - 0.5$	Z_i
1	3.89	0.055	-0.4450	-1.60 (table)
2	4.75	0.166	-0.3333	-0.97
3	4.75	0.277	-0.223	-0.59
4	5.20	0.388	-0.112	-0.29
5	5.78	0.5	0	0
6	5.80	0.611	0.111	0.29
7	6.33	0.722	0.222	0.59
8	7.21	0.833	0.333	0.97
9	7.90	0.944	0.445	1.60

$\bar{x} = 5.7344$

$\sigma = 1.1951$



$z = 0.8368x - 4.734$