Marks Distribution	
Chapter	Marks
Statistical Description of Data	5-7 marks
Measure of Central Tendency	6-8 marks
Measure of Dispersion	5-7 marks

(Theory)
10 marks
+
10 marks
(Numerical)

20 marks

Meaning of Statistics

When defined in Plural sense – Data collection for some specific purpose

When defined in Singular sense – Science or Process of collection, presentation & analysis of data

What topics we are going to cover

- **Collection of Data**

- **Presentation of Data**

- **Present Data Graphically**

What is Statistics

Statistics is derived from –

Latin word 'status', $\longrightarrow$ American

Italian word 'statista' $\longrightarrow$ Art

German word 'statistik' $\longrightarrow$ Strict

French word 'statistique' $\longrightarrow$ etiqute

10 atr. $\longrightarrow$ 2 bar.

Q Statistics is derived

It

× Gres $\longrightarrow$ Status

Application of Statistics

Every field where you need to understand and analyze data is an application of statistics.

That's why we will limit ourselves to only field related to commerce i.e.,

Economics

Business Management

Commerce and Industry

Limitations of Statistics

(1) Study of Quantitative data only: - Statistics studies only such facts as can be expressed in numerical terms. It does not study qualitative phenomena like honesty, friendship, wisdom, health, patriotism, justice, etc.

→ numbers

(2) Study of Aggregates only: - Statistics studies only the aggregates of quantitative facts. It does not study statistical facts relating to any particular.

(3) Homogeneity of Data, an essential Requirement: - To compare data, it is essential that statistics are uniform in quality. Data of diverse qualities and kinds cannot be compared.

→ ek tube ka

(4) Results may Prove to be Wrong: - Future projections of sales, production, price and quantity etc. are possible under a specific set of conditions.

(5) Can be used only by the Experts: - If we do sampling and the rules for random sampling are not strictly adhered to, the conclusion drawn on the basis of these unrepresentative samples would be erroneous.

1. On the nature of data, we can define data as –

Quantitative Data – *numeric data* :- No. of acc. on road, height
  Cardinal Data   weight etc.

Qualitative Data - *honesty, beauty, bravery etc.* → Characterstic which is measurable
                                              → attribute :-Ranking/Scaling

Qualitative information can be used in statistics by converting it to quantitative information by providing a numeric description to the given characteristic. For example – putting it on scale or ranking system

Variable → Vary able } Quantitative data

Discrete Variable:-

When variable taken is countable, then it is called as discrete variable

Example : Money in an account, No. of shares of company

Continuous Variable:-

When variable can obtain any value from the given interval.

Example : Height , weight etc.

179 - 180

180 cm

179.68 cm

179.682 cm

[92 - 93] kg

On the Basis of Source of Data

Primary Data: - Data collected by the investigator for his own purpose, for the first time, from beginning to end, are called primary data.

- These are collected from the source of origin.
- Data, originally collected in the process of investigation are known as primary data.

➢ Interview method

➢ Mailed questionnaire method

➢ Observation method

➢ Questionnaires filled and sent by enumerators.

Interview Method

➢ Direct Interview Method or Personal Interview Method

E.g. – Natural Calamity

➢ Indirect Interview Method

E.g. – Railway accident

➢ Telephonic Interview Method

Fastest interview method with maximum non responses and less accurate

Mailed Questionaire Method

= No. of responses is very low

Google - form

➤ Under this method, questionaires are mailed to the informants. A letter is attached with the questionnaire giving the purpose enquiry. The informant notes the answers against the questionnaires and returns the completed questionnaire to the investigator.

➤ Although a wide area can be covered using the mailed questionnaire method, the number of non-responses is likely to be maximum in this method

Observation Method

:- PT I teacher

Direct information is collected by taking the observation by the observer.

Although this is likely to be the best method for data collection, it is time consuming, laborious and covers only a small area.
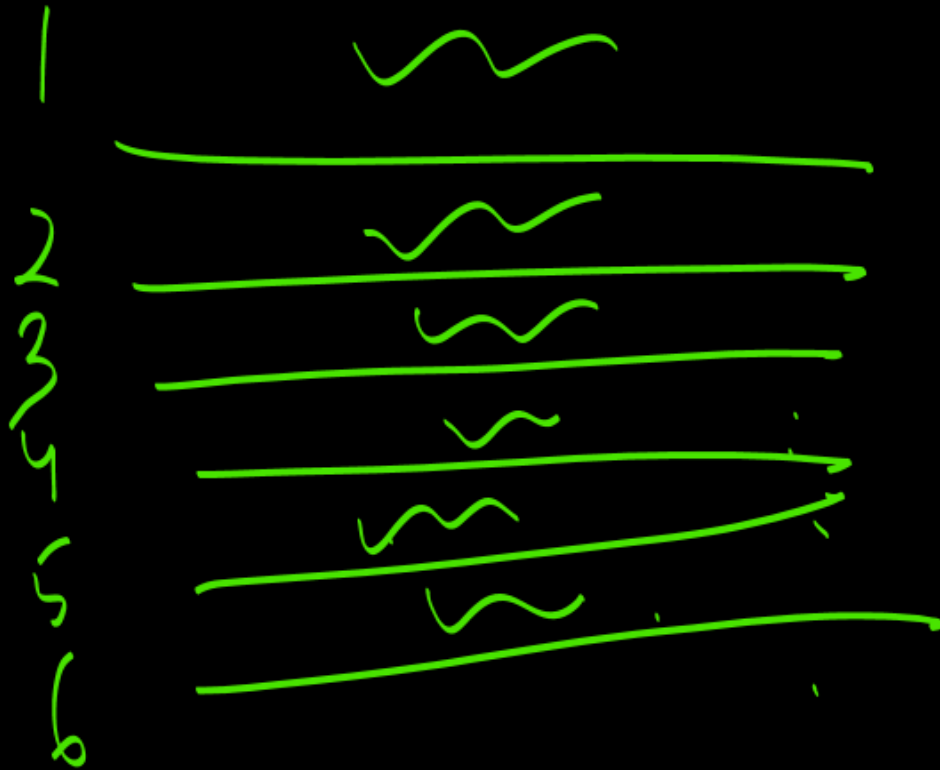
Enumerator's Method → *Aadhar wale Bhai*

↳ *Addhar wale*

➢ Under this method, a questionnaire is prepared according to purpose of enquiry.

➢ To cover the wider range the enumerators are appointed.

This method is very best for wider range but costly in nature

Secondary Data – Secondary data are those which are already in existence, and which have been collected, for some other purpose.

These data are, therefore, called second-hand data. Because data have already been collected by somebody else, these are available in the form of published or unpublished, reports

School ka Principal → PT

Healthy → BMI → Height / Weight

Unhealthy → BMI
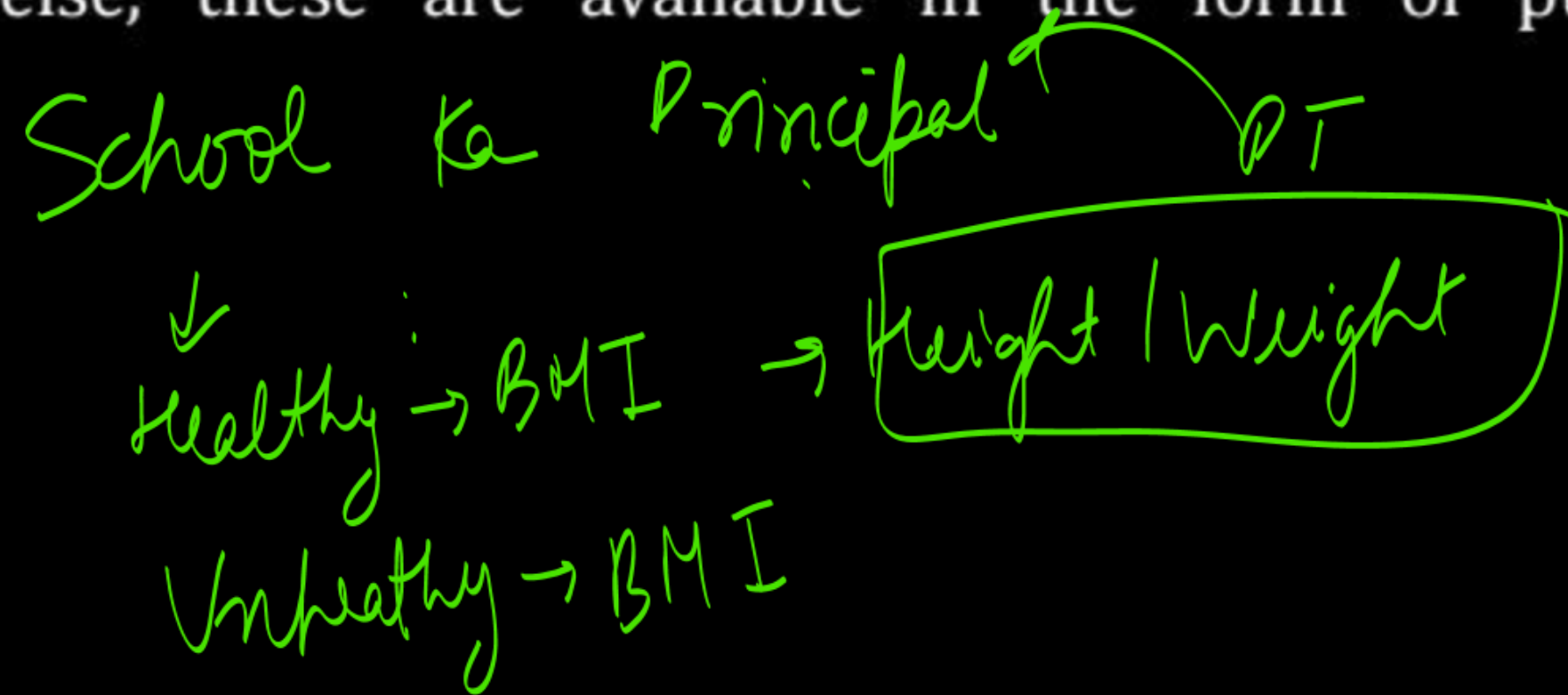
Sources of Secondary Data

There are many sources of getting secondary data. Some important sources are listed below:

- International sources like WHO, ILO, IMF, World Bank etc.

- Government sources like Statistical Abstract by CSO, Indian Agricultural Statistics by the Ministry of Food and Agriculture and so on.

- Private and quasi-government sources like ISI, ICAR, NCERT etc.

- Unpublished sources of various research institutes, researchers etc

Scruitny of Data → Homogenous

Checking Accuracy and Consistency

Accuracy :- Relatable multiple
data
Date of Birth ∽ Age

Internal Consitency of Data :When two or more related data are present , cross checking can happen

Presentation of Data

➢ Once the data are collected and verified for their homogeneity and consistency, we need to present them in a neat and condensed form highlighting the essential features of the data.

➢ Any statistical analysis is dependent on a proper presentation of the data under consideration.

➢ An important method of organization of data is to distribute these into different classes on the basis of their characteristics. This process is called classification of data

- It involves conversion of raw data into groups in a manner such that some meaningful conclusions can be drawn out of them.

- Basically, Classification of Data can be defined as the process of arranging data on the basis of characteristics into the number of groups or classes according to the similarity of observations

$0 - 100$

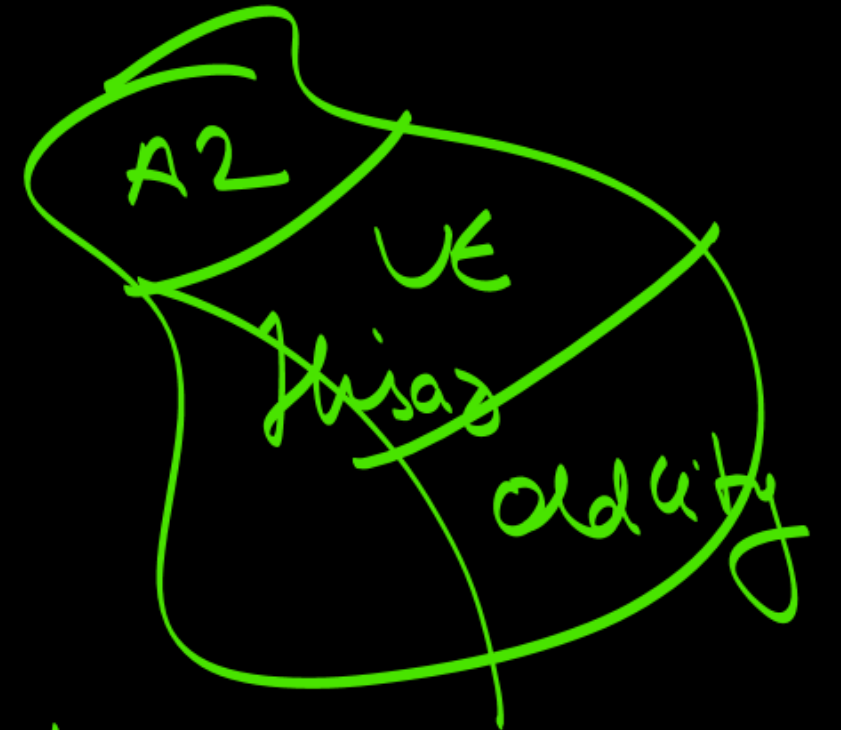|	
0 - 20	5
20 - 40	6
40 - 60	8
60 - 80	10
80 - 100	3

1 Class
2nd Class

Data may be Classified as

(i)   Data Chronological or Temporal or Time Series: -

Jan — March

April — June

July — Aug

Sep — Dec

A2

UE

Hisar

Old city

(ii)   Geographical or Spatial Series Data: -.

AN

UE

Old city

Sector

Data may be classified as

(iii) Qualitative or Ordinal Data: -

$$\frac{Badmash}{Kam\ Badmash}$$
$$\frac{Ache\ Bache}{Bhaut\ ache}$$

(iv) Quantitative or Cardinal Data: -

$$\frac{0-20}{20-40}$$
$$\frac{40-60}{60-80}$$

Frequency Data

The frequency of a particular data value is the number of times the data value occurs and data in which we can count frequency is called as frequency data

Example:

	No. of things
Bangla	1000 ← Frequency
Swiss Watch	20000
Sports Car	100

Non-Frequency Data

Data where the identity of the each of the individual values has to be kept in view are called Non-frequency Type data.
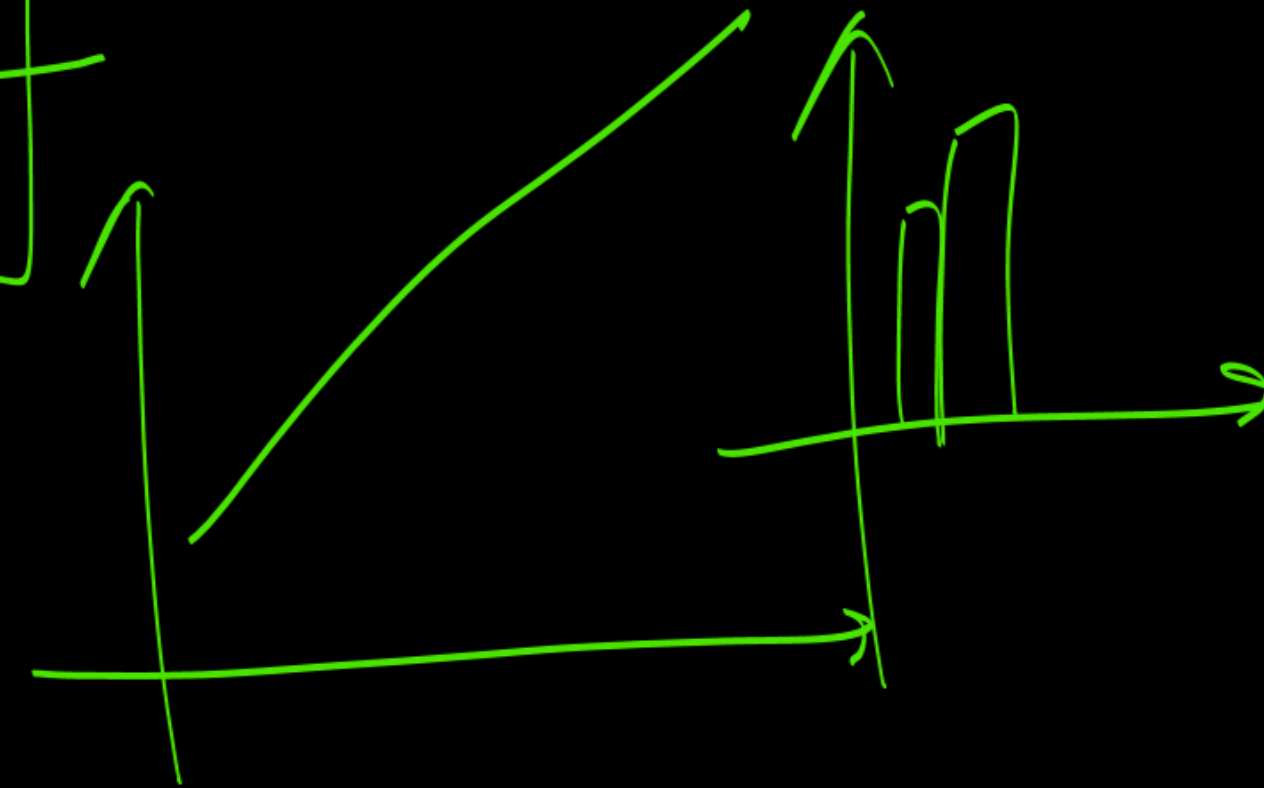
Example:

Chair	ı
Car	ı
Watch	ı

There are Generally Three Forms of Presentation of Data

❑ **Textual or Descriptive Presentation**

❑ **Tabular Presentation**

❑ **Diagrammatic Representation of Data**

Textual or Descriptive Presentation

In the textual presentation, data is described by paragraph of text. This method is used in most of the official reports, where the activities, plans or programmers of a project are described in words, inserting relevant facts and figures in between them. When the quantity of data in not too large, this form of presentation is most suitable.

Example:

The Company could not achieve collection efficiency targets during the last five years ending March 2018 as prescribed by HERC and resultantly recoverable amount had increased from ₹ 4,460.18 crore in March 2014 to ₹ 7,332.70 crore in March 2018.

(Paragraph 2.1.10.1)

Chapter III discusses Transaction audit observations which highlight deficiencies in the management of State Government Companies of power sector, which had serious financial implications. Important findings are as under:

Uttar Haryana Bijli Vitran Nigam Limited

- The Company incurred extra expenditure of ₹ 5.34 crore on purchase of transformer oil by resorting to limited tender enquiry instead of open tenders. The Company could not utilise the inventory of ₹ 198.54 crore due to delay in receiving quality test reports from NABL empanelled laboratories. As on 31 March 2018, shortages of ₹ 1.73 crore were pending investigation.

(Paragraph 3.1)

Haryana Power Purchase Centre

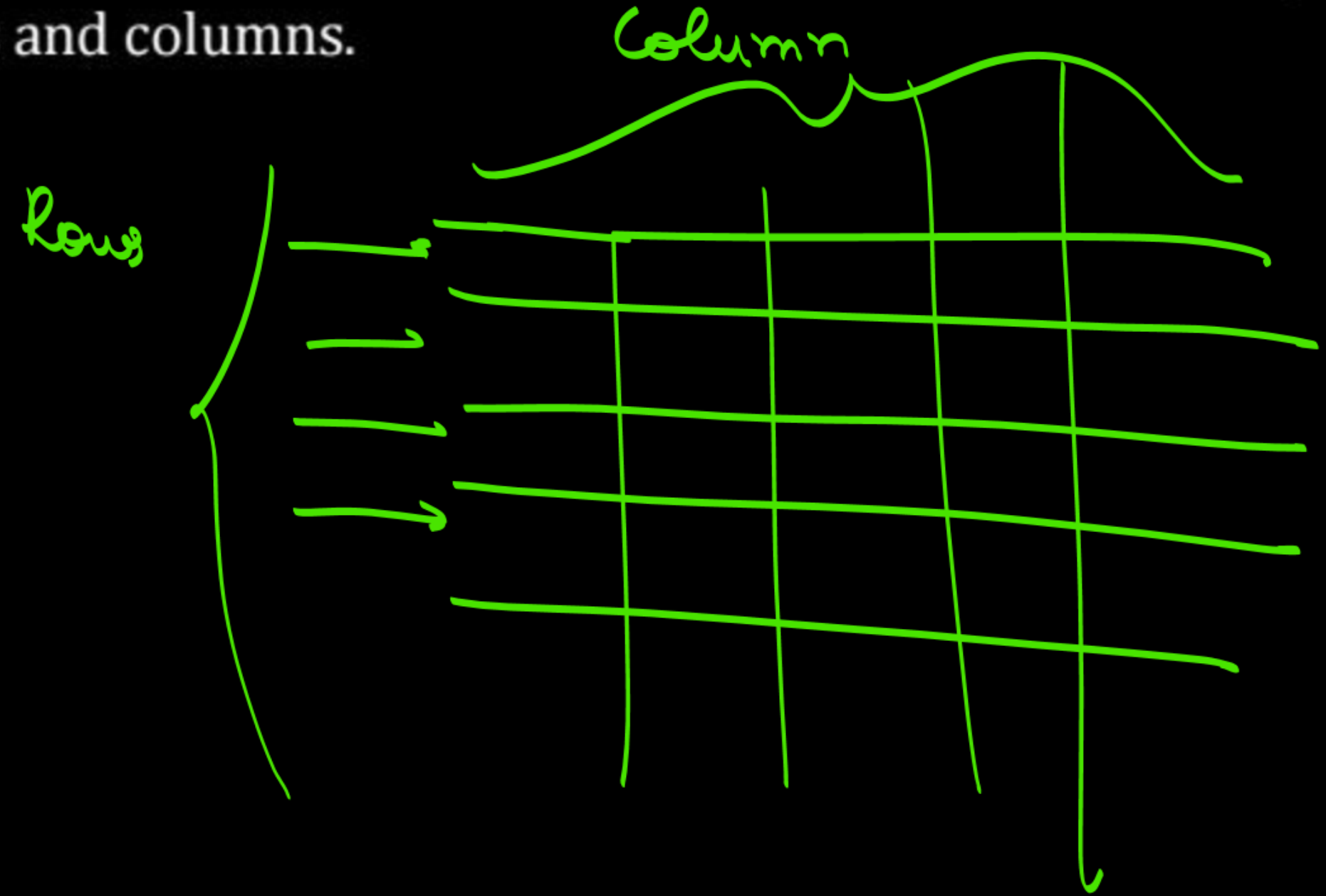✓ **Unpopular**

✓ **Difficult to interpret**

✓ **Time consuming**

Tabular Presentation

Data is presented in rows and columns. A statistical table is a systematic presentation of data is rows and columns.

Left most part
Defining the rows

FORMAT OF TABLE

Table Number: _____

Title: _____

(Head Note, if any)

→ Boxhead

Stub (Row Heading)	Caption (Column Heading)				Total (Rows)
	Sub-head		Sub-head		
	Column-head	Column-head	Column-head	Column-head	
Stub E. 1 Stub Entries (Row Entries)					
St. En 2				BODY	
St Gnt. 3			← BODY →		
. . .					
—					
Total (Columns)					

Source Note:
Footnote: **Kg → Kilogram**

Diagrammatic Representation of Data

Diagrammatic presentation of data is another useful method of presenting the data in a compact form. There are various kinds of diagrams in common use. But we will have discussion regarding: -

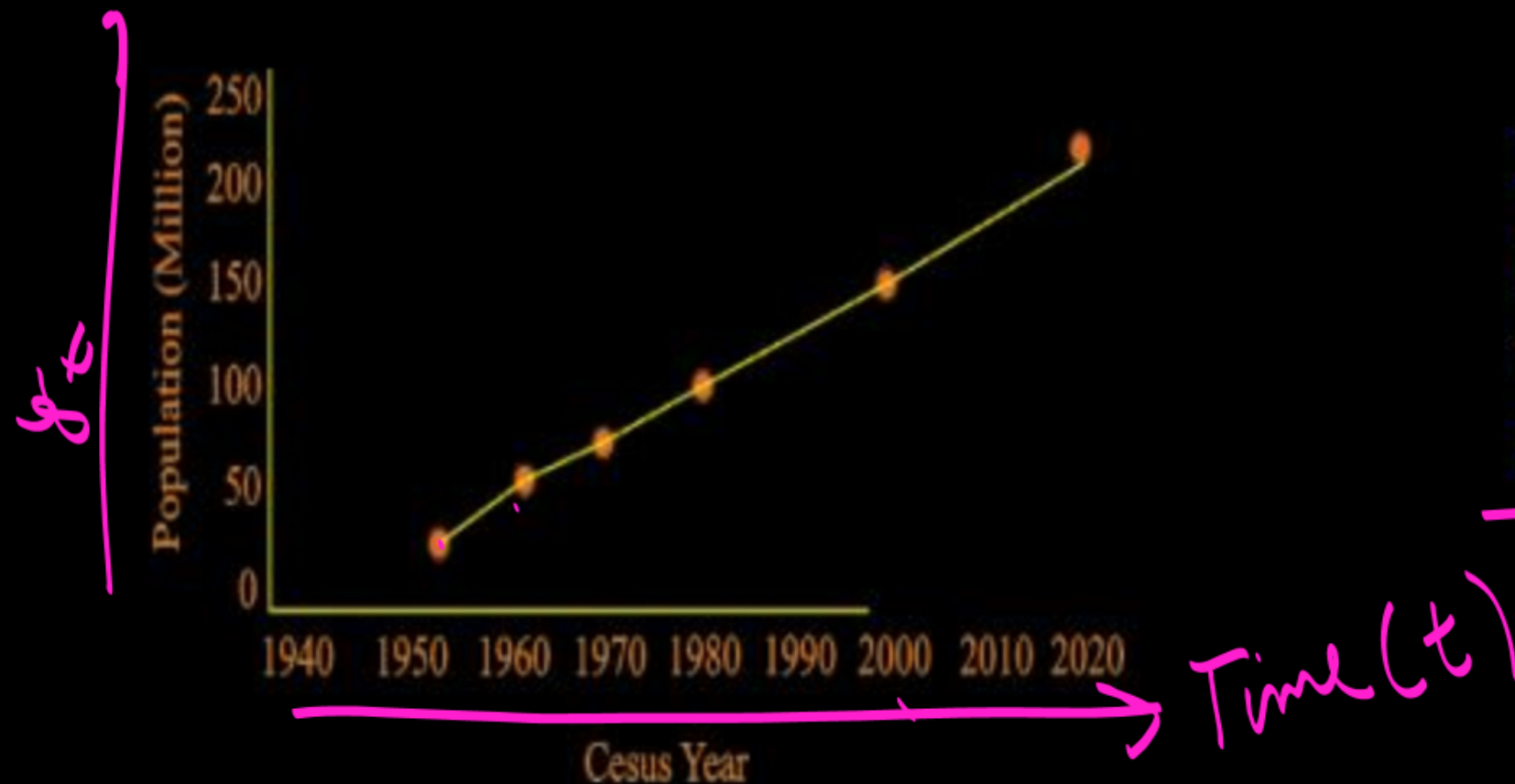(i) Line Diagram or Historiagram

(ii) Bar Diagram

(iii) Pie Chart

Whenever data vary with time, we use line diagram which is also known as Historiagram. (Generally, used in time series)

- In a simple line diagram, we plot each pair of values of (t, y_t), y_t representing the time series at the time point t in the $t - y_t$ plane. snd then the points are joined

If the fluctuation in data is high, we present the data in form of logarithm vs time known as Log Chart and Ratio Chart
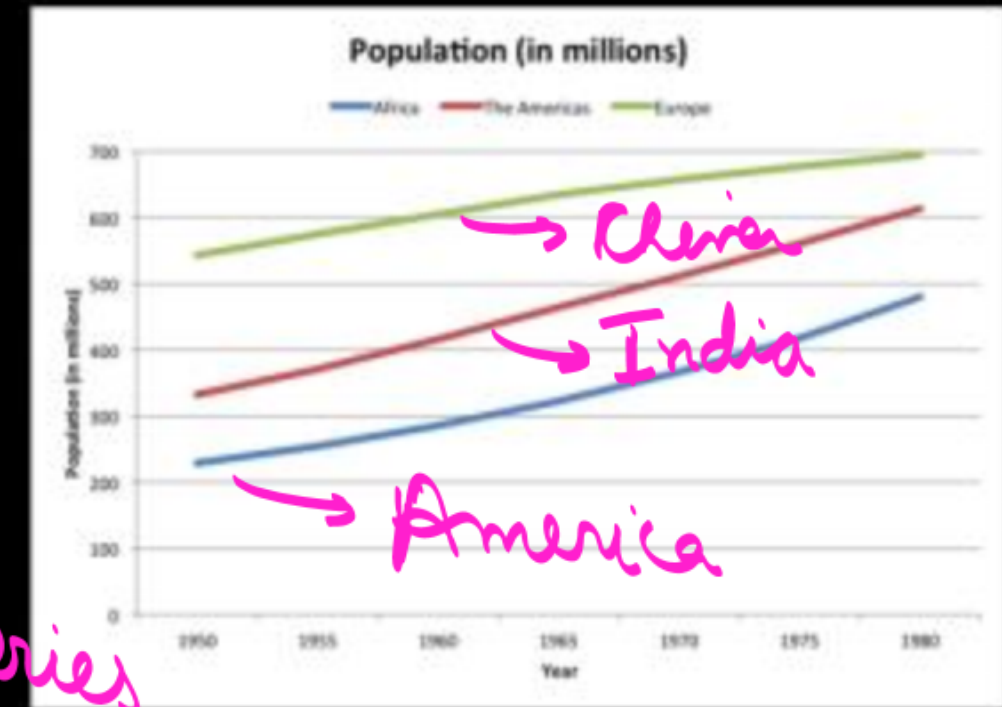
logarithm

$\log_{10} 10 = 1$

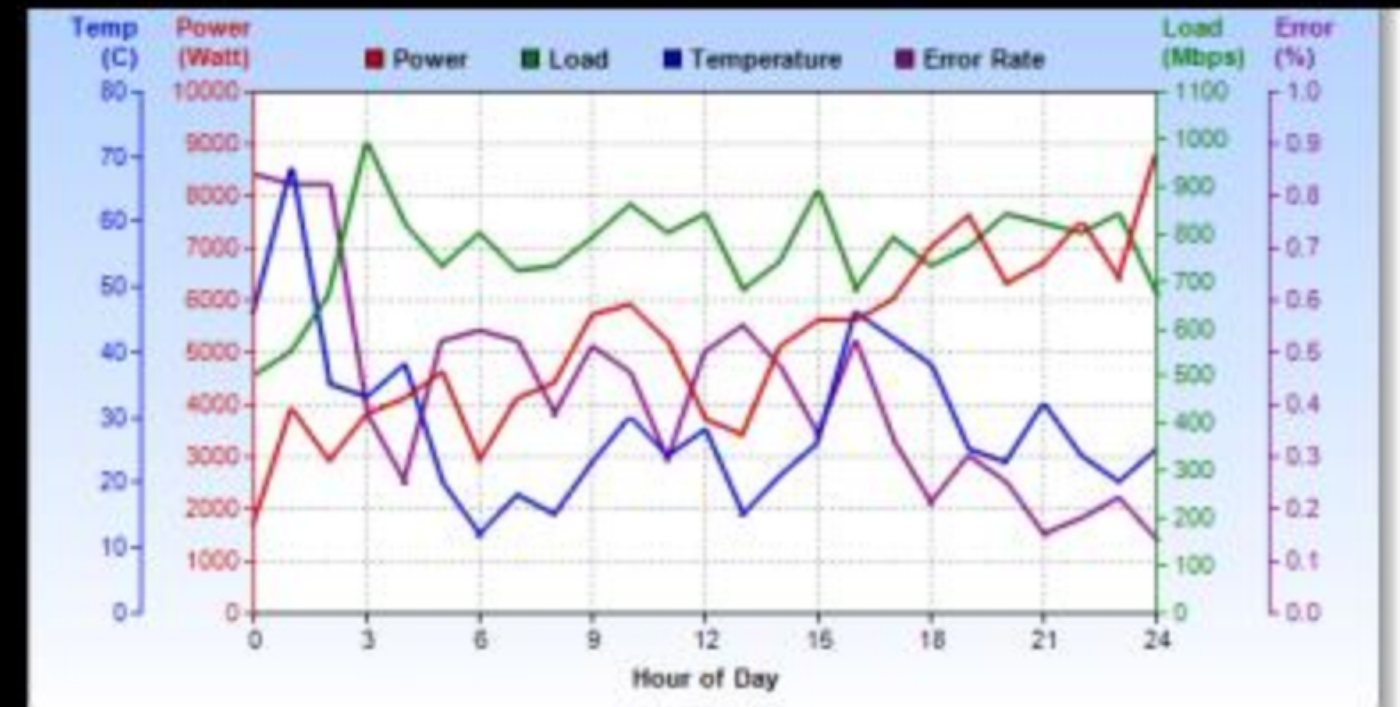|	10
3	1000
5	100000
7	10000000
11	100000000000

Multiple line chart and Multiple Axis Chart

- If we have to represent multiple data (similar data in same unit) varying with time, we can use Multiple line chart.

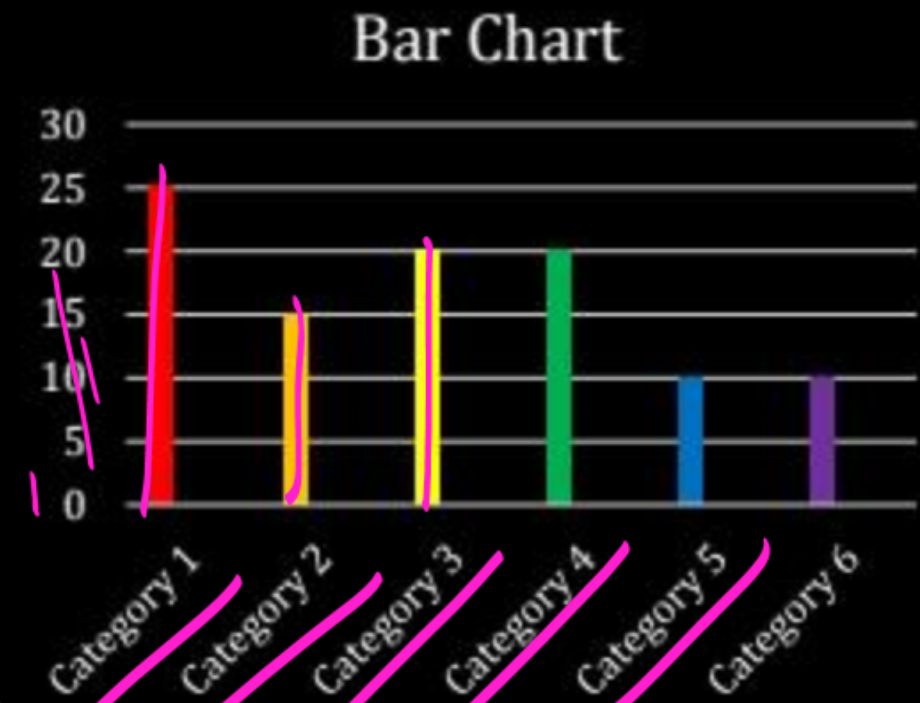*relatable data in same unit present in time-series*

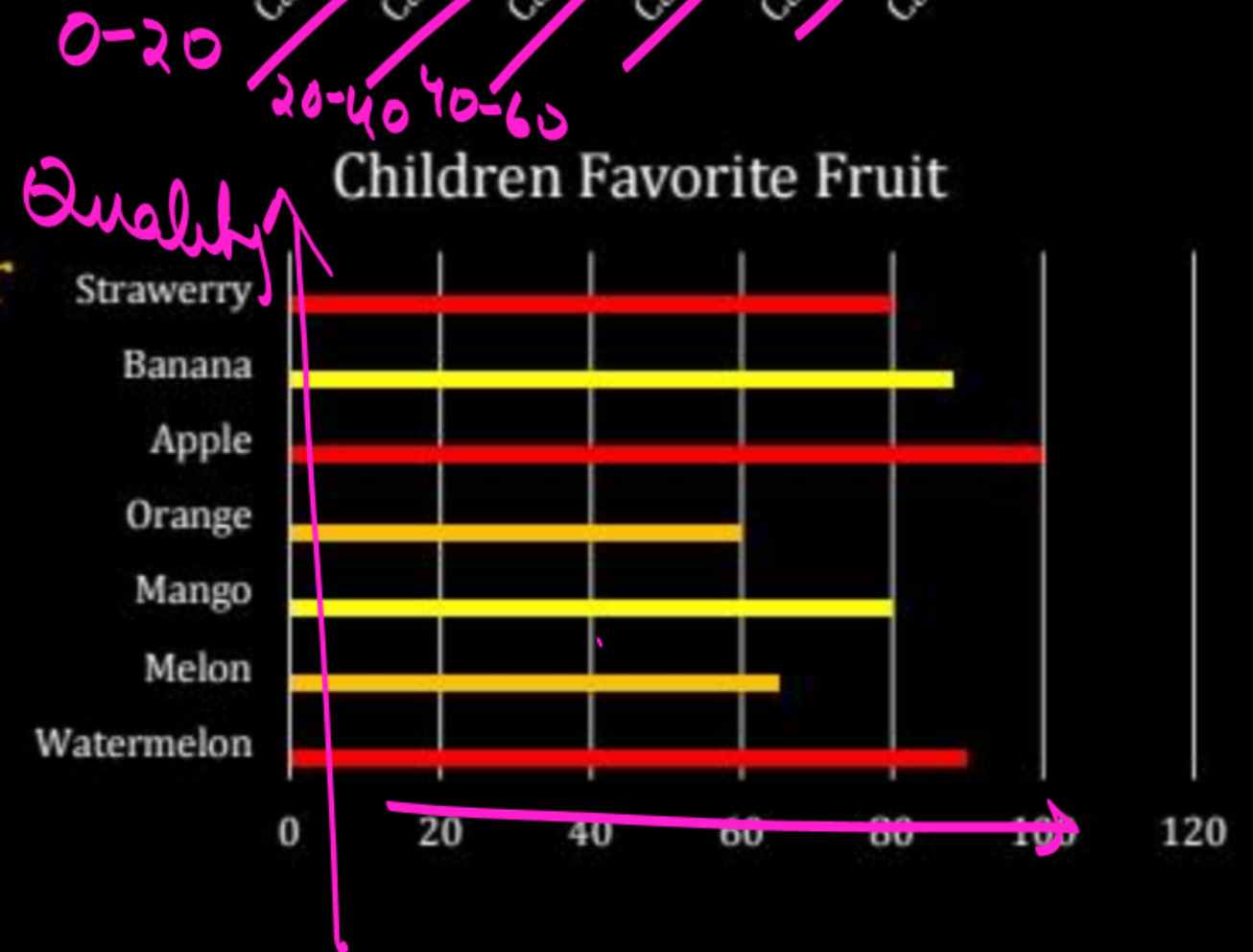- If we have to represent multiple data (with different unit) varying with time, we can use Multiple axis chart.

Bar Diagram

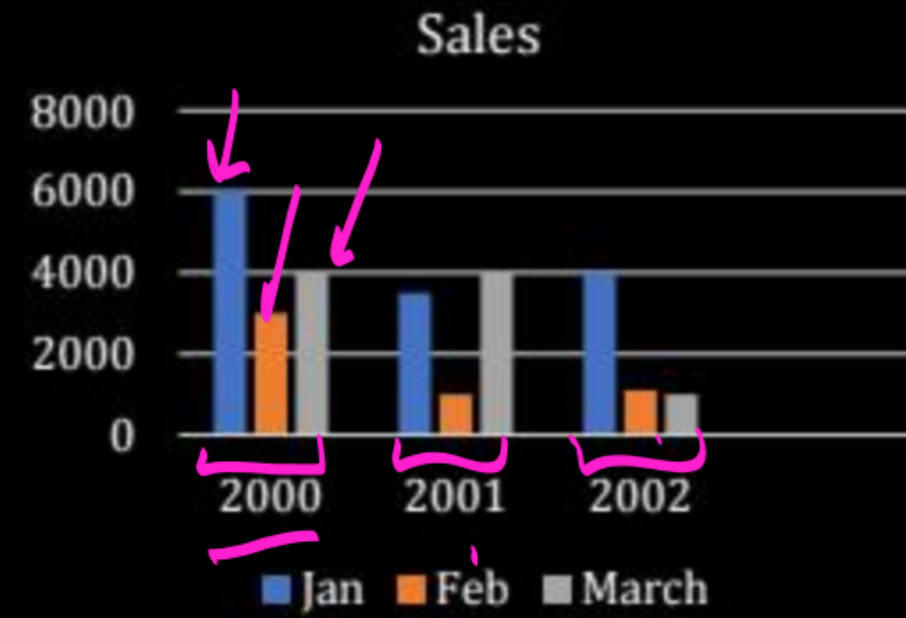**Vertical Bar Diagram** – Used for Quantitative and Time series Data *or Temporal Data*

Horizontal Bar Diagram – Used for Qualitative or Geographical Data *or Spatial Data*
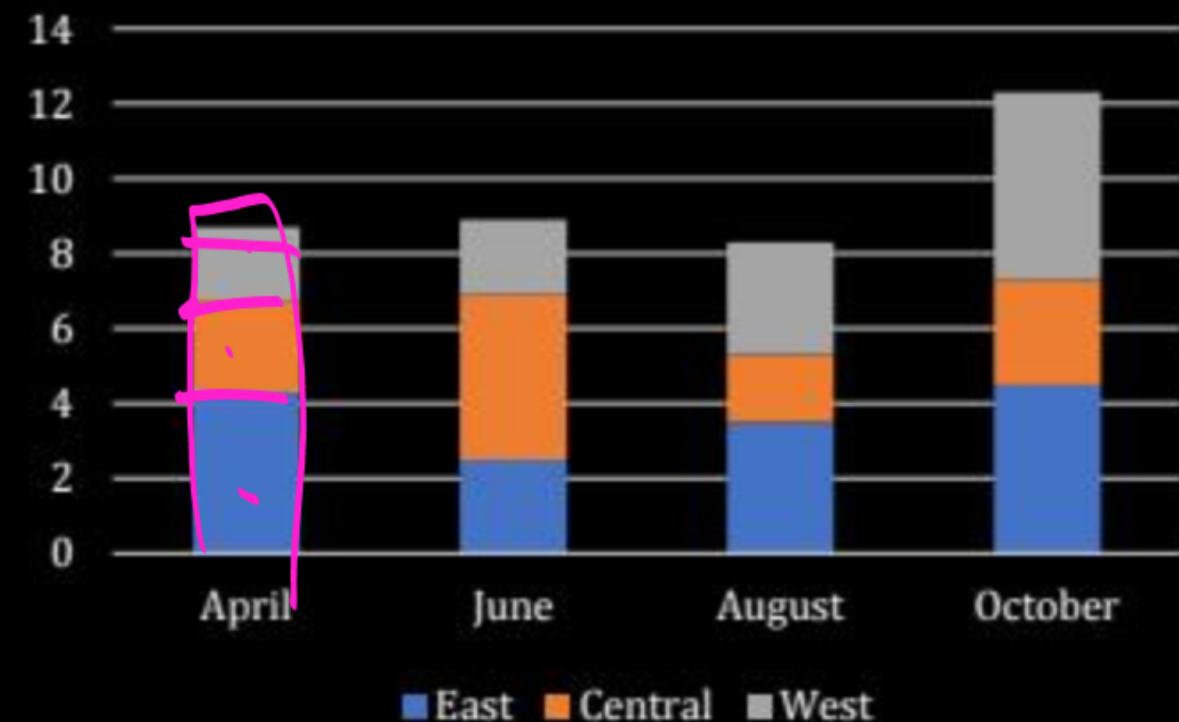
0-20   20-40 40-60

Quality

Bar Chart

Children Favorite Fruit

Bar Diagram

Multiple or Grouped Bar diagrams –

For Comparable series

Component or sub-divided Bar diagrams –

Data Divided into Multiple Component

Bar Diagram

Divided Bar Diagram / % Bar diagram

For relative Comparision with whole

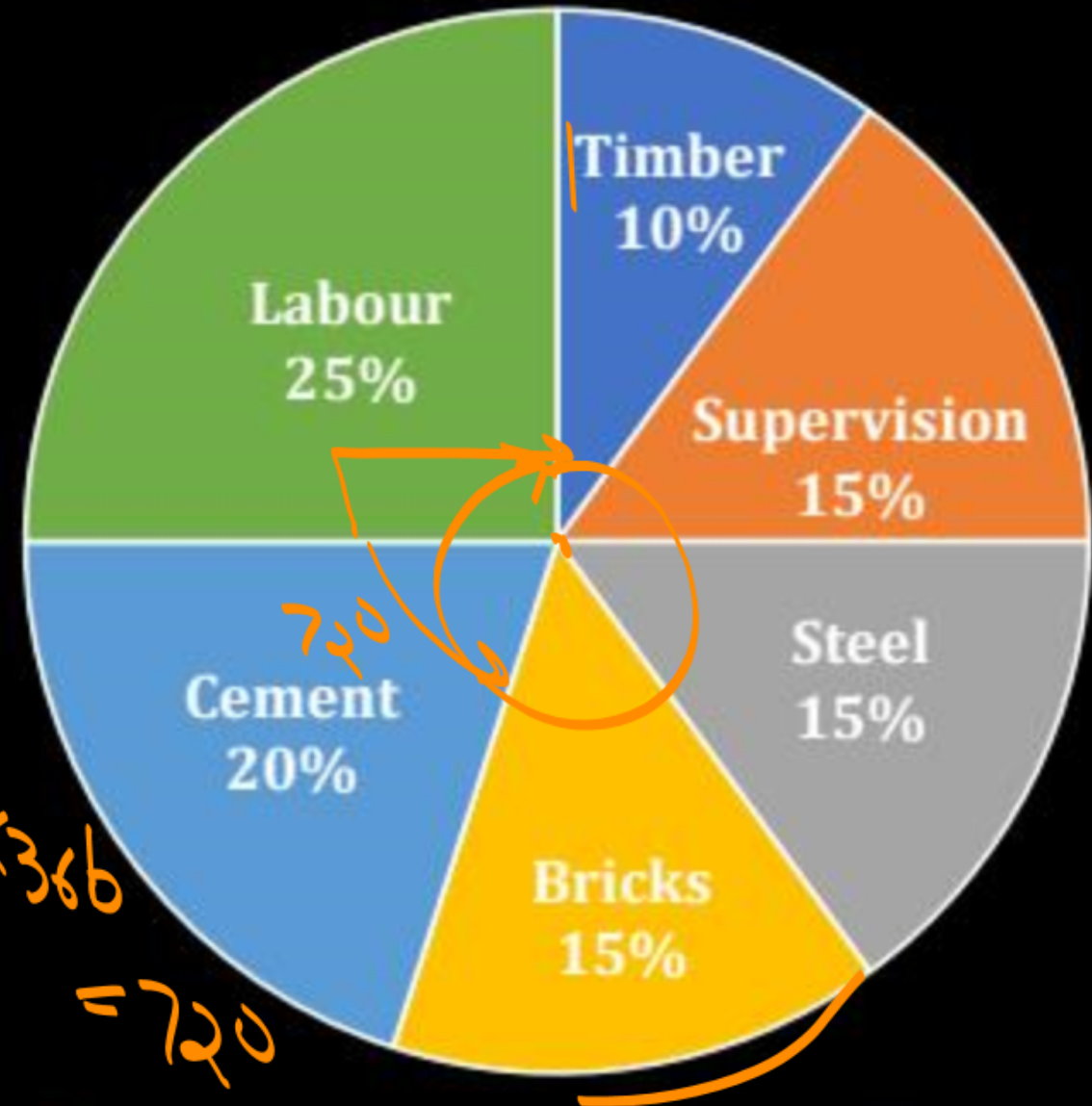As well as comparision between different groups.

Weekly Product Sales

Pie Diagram

Total :- 25 + 10 + 15 + 15 + 15 + 20 = 100

Formula for Angle Calculation :-

Angle – (Segment Value * 360) / Total Value

Used for presenting a part with a whole

$$\text{Angle} = \frac{25\%}{100\%} \times 360 = 900 \quad \Bigg| \quad \frac{20 \times 360}{100} = 720$$

Cost of Construction of House

- Timber 10%
- Supervision 15%
- Labour 25%
- Steel 15%
- Cement 20%
- Bricks 15%

72°

□ Timber □ Supervision □ Steel □ Bricks □ Cement □ Labour

The production of Wheat by different States of India are as shown below: -

2000

State	Production
Haryana	20
Punjab	25
Uttarakhand	25
Up	20
Others	10%
Total	100%

$$\frac{20}{100} \times 360 = 72°$$

$$\frac{25}{100} \times 36\emptyset = 90°$$

$$\frac{25}{100} \times 360° = 90°$$

$$\frac{20}{100} \times 360° = 72°$$

$$\frac{10}{100} \times 360 = 36°$$
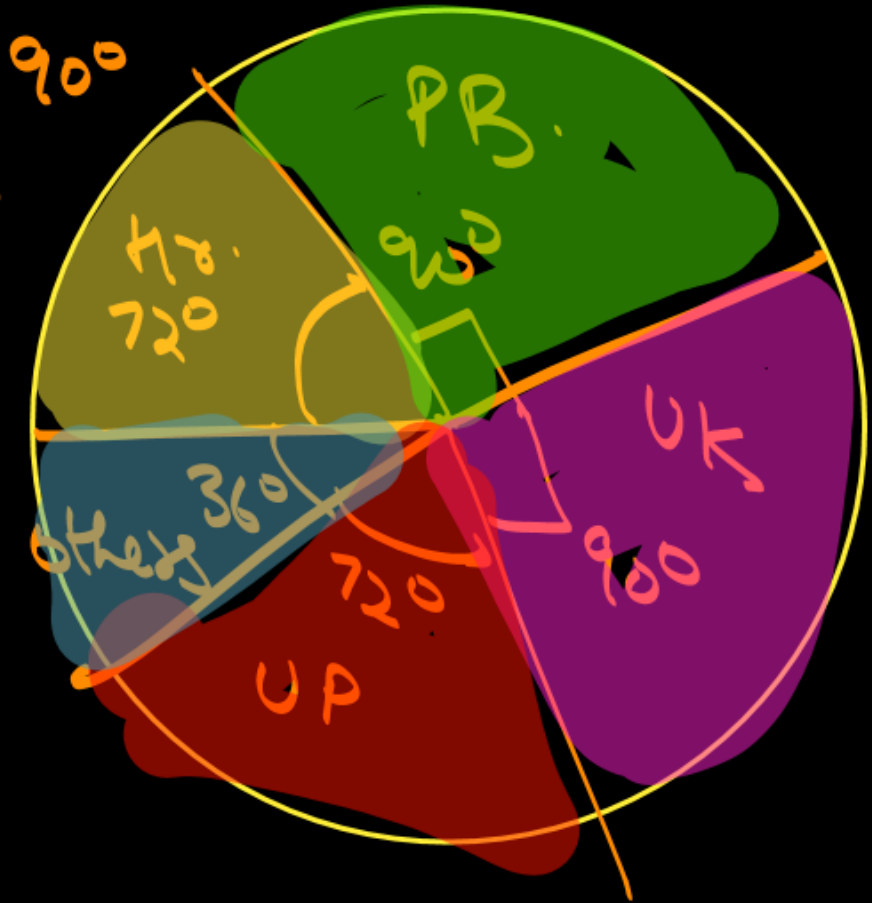
Draw suitable diagram to represent the information

The mode of presentation of data are

A Textual, tabulation and diagrammatic

B Tabular, internal and external

C Textual, tabular and internal

D Tabular, textual and external.

The best method of presentation of data is

(A) Textual

(B) Tabular $\rightarrow$ Best | Expert

(C) Diagrammatic $\rightarrow$ Hidden Trend | Noob

(D) (b) and (c).

For tabulation, 'caption' is

A The upper part of the table

B The lower part of the table

C The main part of the table

D The upper part of a table that describes the column and sub-column.
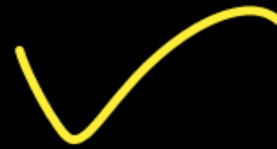
'Stub' of a table is the

(A) Left part of the table describing the columns

(B) Right part of the table describing the columns

(C) Right part of the table describing the rows

(D) Left part of the table describing the rows.

The unit of measurement in tabulation is shown in

A Box head

B Body

C Caption

D Stub

Which of the following statements is untrue for tabulation?

A Statistical analysis of data requires tabulation

B It facilitates comparison between rows and not columns

C Complicated data can be presented

D Diagrammatic representation of data requires tabulation.

Diagramatic presen

↓

Table
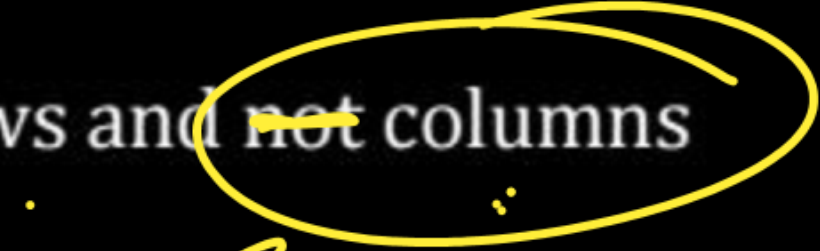
Hidden trend, if any, in the data can be noticed in

(A) Textual presentation

(B) Tabulation

(C) Diagrammatic representation

(D) All these
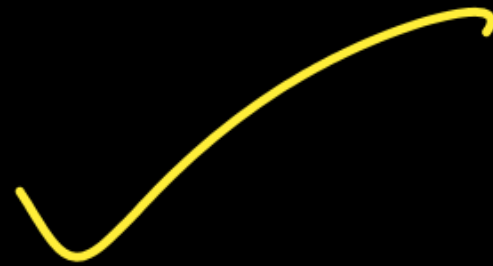
Multiple line chart is applied for

A Showing multiple charts

B Two or more related time series when the variables are expressed in the same unit → Multiple line chart

C Two or more related time series when the variables are expressed in different unit → Multiple axis chart

D Multiple variations in the time series.

Horizontal bar diagram is used for

A Qualitative data

B Data varying over time → Time Series

C Data varying over space

D (A) or (C)

→ Geographical

Divided bar chart is considered for

A Comparing different components of a variable

B The relation of different components to the table

C (A) or (B)

D (A) and (B)

Pie-diagram is used for

A Comparing different components and their relation to the total

B Representing qualitative data in a circle

C Representing quantitative data in circle

D (b) or (c)
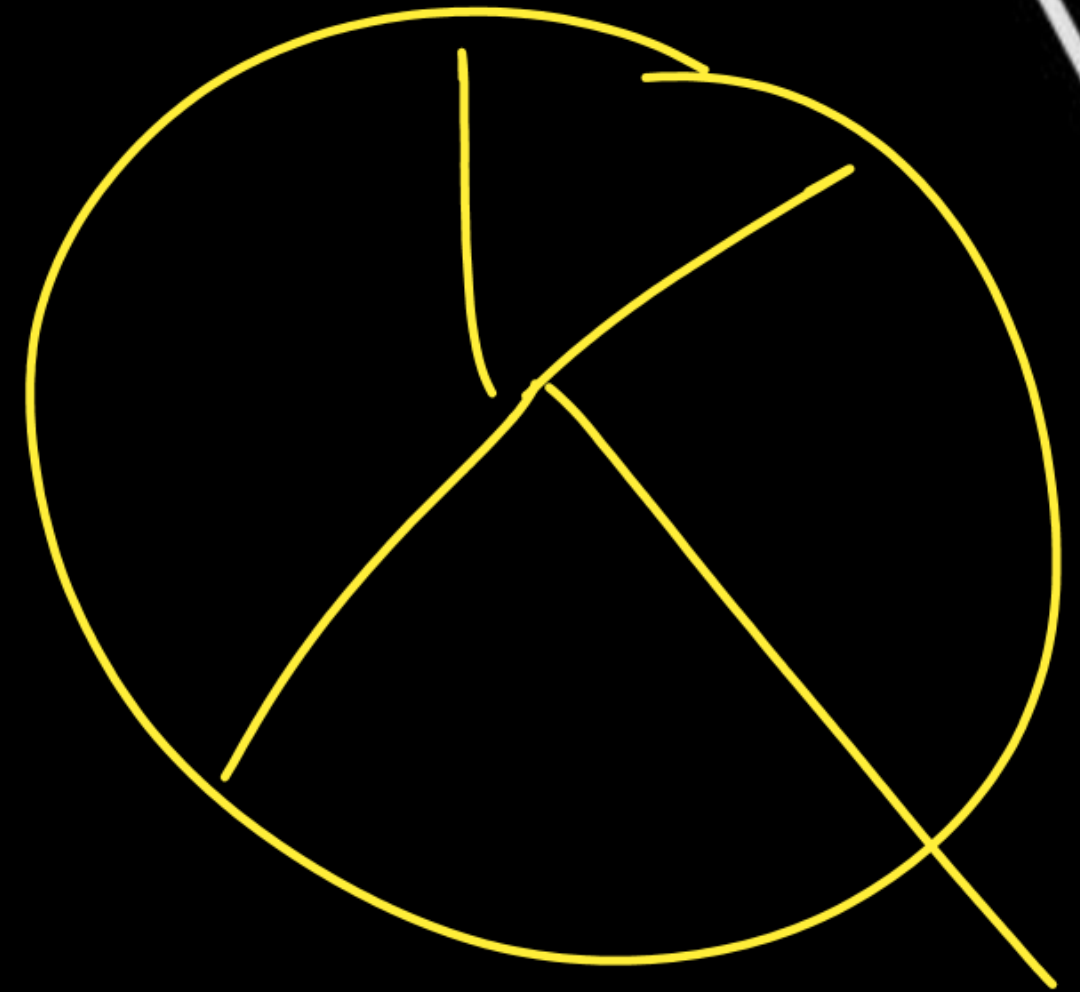
Frequency Distribution *(Counting)*

- 1, 2, 3, 4, 6, 9, 9, 8, 5, 1, 1, 9, 9, 0, 6, 9.

  The frequency of number 9 is: **5**

- Frequency Distribution can be defined as the tabular representation of data in which we distribute total frequency to the number of classes.

Ungrouped freq dist
→ Discrete Series

Grouped freq Dist
↳ Discrete series
↳ Continuous series

Ungrouped Frequency Distribution

- When we draw a tabulation for the frequency of data where data is in the form of discrete variable.

Example:

Let's say you survey a number of households and find out how many pets they own. The results are 3, 0, 1, 4, 4, 1, 2, 0, 2, 2, 0, 2, 0, 1, 3, 1, 2, 1, 1, 3. Let's distribute data in frequency table.

No. of Pets	Tally Marking	No. of Households
0		4
1		6
2		5
3		3
4		2

Grouped Frequency Distribution

- When we draw a tabulation for the frequency of data where data is in the form of discrete variable.

Example:

21 yr

21 yr 4m

21 yr 4m 3d

21 yr 4m 3d 20hr

21-22

Grouped

Count / Freq.

Age Group	Tally Marks	No. of People
0 – 10	卌	5
10 – 20	IIII	4
20 – 30	III	3
30 – 40	IIII	4
40 – 50	卌	6
50 – 60	II	2
60 – 70	卌 I	6
		30

$1-5 \rightarrow$ **Limits of class**

$L.C.L$ $U.C.L$

Discrete

Road accident

No. of accident	No. of Roads
1 - 5	4
6 - 10	3
11 - 15	3
16 - 20	5
21 - 25	10
	25

Grouped freq

$L.C.B$

$1-5$
$6-10$ } **Class Interval**

Continuous

Age

$0-10 \rightarrow$ **Boundaries**

$U.C.B$

$0-10$
$10-20$ } **Class Interval**

Age Group	No. of people
0 - 10	5
10 - 20	4
20 - 30	3
30 - 40	4
40 - 50	6
50 - 60	2
60 - 70	6
	30

Boundary

U.C.B of pv. class
= L.C.B of next class

- **Class Limit:** Corresponding to a class interval, the class limits may be defined as the minimum value and the maximum value of the class interval may contain. The minimum value is known as _L.C.L_____ and the maximum value is known as the _U.C.L_____

Example :

Lower Class Limit

$$158 - 161$$

Upper Class Limit
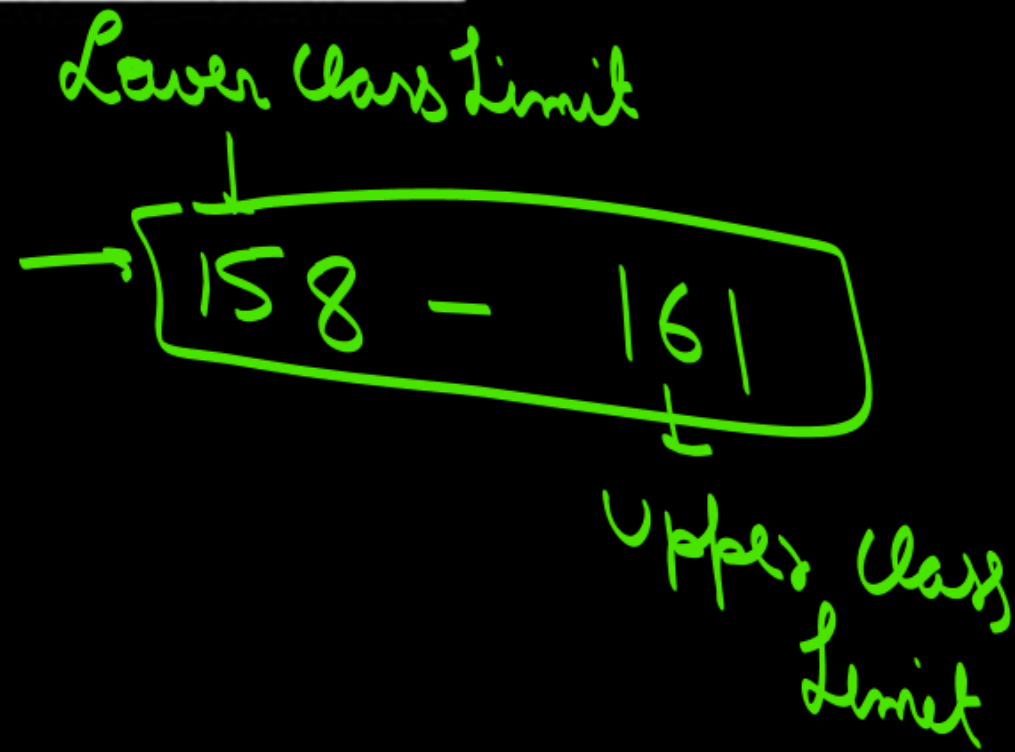
Class – Interval	Frequency
150-153	7
154-157	7
158-161	15
162-165	10
166-169	5
170-173	6

Some Important Terminologies

➤ **Class boundary:** Class boundaries may be defined as the actual class limit of a class interval.

Continuous variable

For overlapping classification or mutually exclusive classification

Example:

$\rightarrow$ *Exclude U.C.B while taking the value.*

For non-overlapping or mutually inclusive classification

Example: *discrete* $\rightarrow$ $L.C.B = L.C.L - \dfrac{gap}{2}$. $U.C.B = U.C.L + \dfrac{gap}{2}$

➤ **Please note that for overlapping class interval:** Class boundaries and Class limits are same.

$$\frac{gap}{2} \qquad \frac{1}{2} = 0.5$$

Find Upper Class Boundary And Lower Class Boundary

Class Limits

Class – Interval	Frequency
150-153	7
154-157	7
158-161	15
162-165	10
166-169	5
170-173	6

Mid

$$\begin{array}{c} L.C.B \\ = L.C.L - 0.5 \end{array} \qquad \begin{array}{c} U.C.B = \\ U.C.L + 0.5 \end{array}$$

151.5	149.5 – 153.5
155.5	153.5 – 157.5
159.5	157.5 – 161.5
163.5	161.5 – 165.5
167.5	165.5 – 169.5
171.5	169.5 – 173.5

L.C.B are less than Lower Class Limit of interval

U.C.B are upper than U.C.L of an interval

$C.L = 173.5 - 169.5 = 4$

$C.L = 173 - 170 = 3$

→ To represent a single value of an interval

Class Mid-Point: The class midpoint (or class mark) is a specific point in the center of the class interval in a frequency distribution table.

Mid-Point =
$$\frac{U.C.L + L.C.L}{2}$$

Example:

or $\dfrac{U.C.B + L.C.B}{2}$

Mid Pt

$\dfrac{150+153}{2} = 151.5$

155.2

159.5

163.5

167.5

171.5

Class – Interval	Frequency
150-153	7
154-157	7
158-161	15
162-165	10
166-169	5
170-173	6

For the non – overlapping classes 0 -19, 20-39, 40-59 the class mark of the class 0 -19 is

A 0

B 19

C 9.5

D None

$$0 - 19$$

$$I$$

Mid pt $\rightarrow \dfrac{0+19}{2} = 9.5$

- **Width or Size of Class Interval:** The class width is the difference between the upper-class boundary to the lower-class boundary of consecutive classes

Example:

Size of C.I

or Class length $= U.C.B - L.C.B$

$= 40 - 30 = 10$

Class – Interval	Frequency
0 – 10	4
10 – 20	8
20 – 30	13
30 – 40	12
40 – 50	6

Cumulative Frequency

- Cumulative frequency is defined as a running total of frequencies.

Cumulative frequency

More Than (L.C.B)

Less Than (U.C.B)
(By default)

Example:

Class	Frequency	(L.C.B) More than	(U.C.B) Less than
0-10	4	43	4 → M.C.B → 4
10-20	8	39 → C.B 10	12 → 39+12
20-30	13	31	25 → C.B 20
30-40	12	18	37
40-50	6	6	43

Total = 43

For a C.B,

C.T.C.F + M.T.C.F

Class boundary	More than	Less than
→ 0	43	0
→ 10	39	4
→ 20	31	12 → 31+12 = 43
→ 30	18	25
→ 40	6	37
50	0	43

➤ Frequency Density of Class Interval-

$$= \frac{Class\ frequency}{C.L.}$$

$C.L = U.C.B - L.C.B = 10 - 0 = 10$

➤ Relative Frequency or % Frequency of Class Interval

$$= \frac{Class\ frequency}{Total\ Frequency} \times 100\%$$

Class	Frequency	Frequency density	Relative frequency |%.
0-10	4	$\frac{4}{10} = 0.4$	$\frac{4}{43} \times 100\% = 9.3\%$
10-20	8	$\frac{8}{10} = 0.8$	$\frac{8}{43} \times 100\% = 18.6$
20-30	13	$\frac{13}{10} = 1.3$	$\frac{13}{43} \times 100\% = 30.23$
30-40	12	$\frac{12}{10} = 1.2$	$\frac{12}{43} \times 100\% = 27.91$
40-50	6	$\frac{6}{10} = 0.6$	$\frac{6}{43} \times 100\% = 13.95\%$

Total = 43

Sum = 100%

HISTOGRAM

- The areas of rectangle are proportional to the frequencies.
- We can find the mode from Histogram

→ Class interval Freq are of equal length

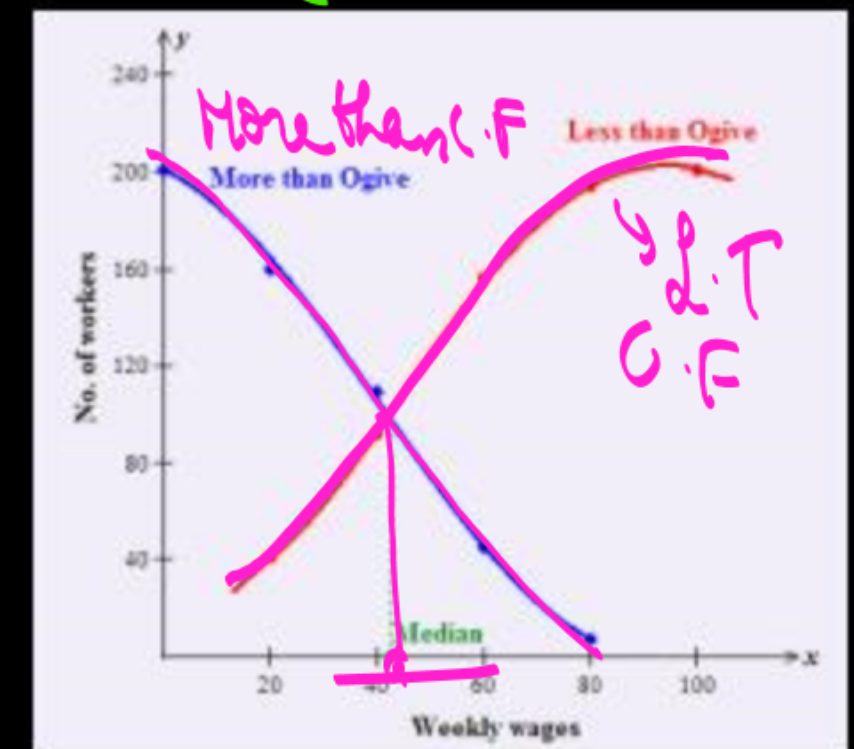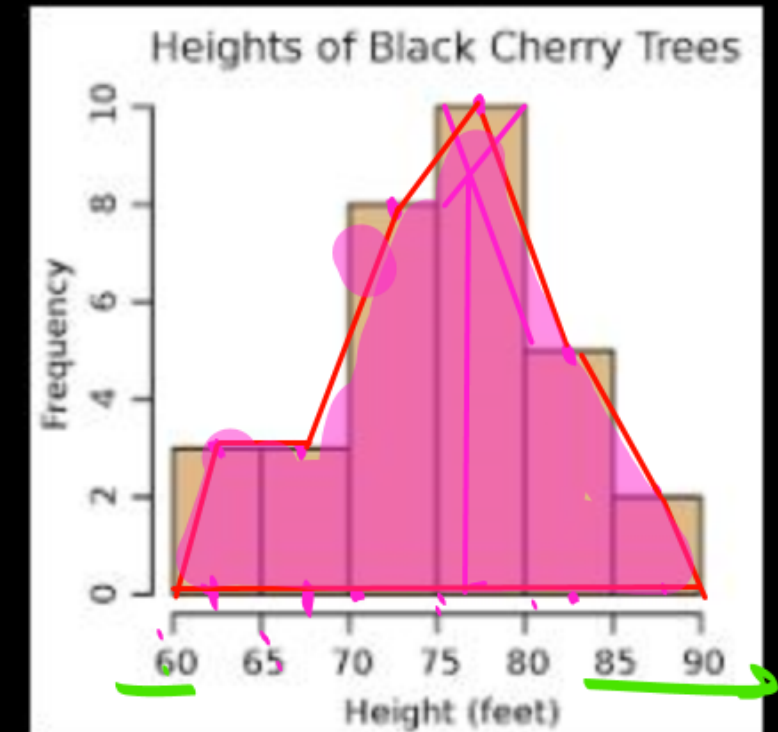Heights of Black Cherry Trees
Frequency / Height (feet)

FREQUENCY POLYGON: -

- We can draw frequency polygon by plotting the (x_i, f_i) and then joining it with the line.

mid point → freq.

OGIVES OR CUMULATIVE FREQUENCY GRAPH

- Ogives help us to find out Median.
- Ogives help us to find out Quartiles

More than C.F   Less than Ogive
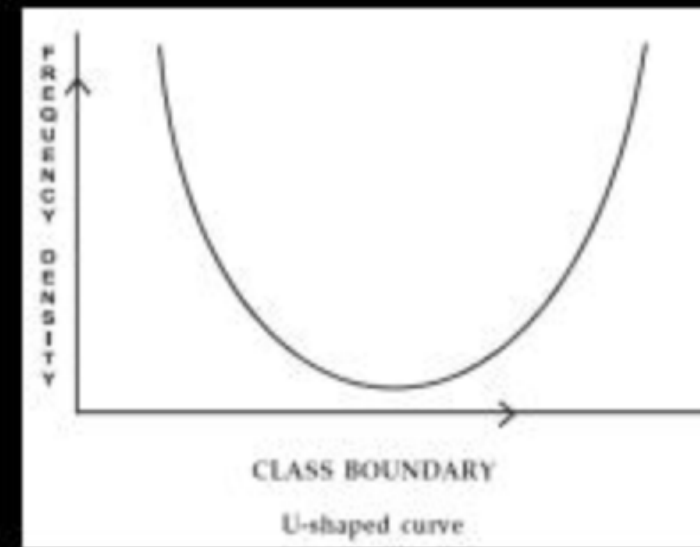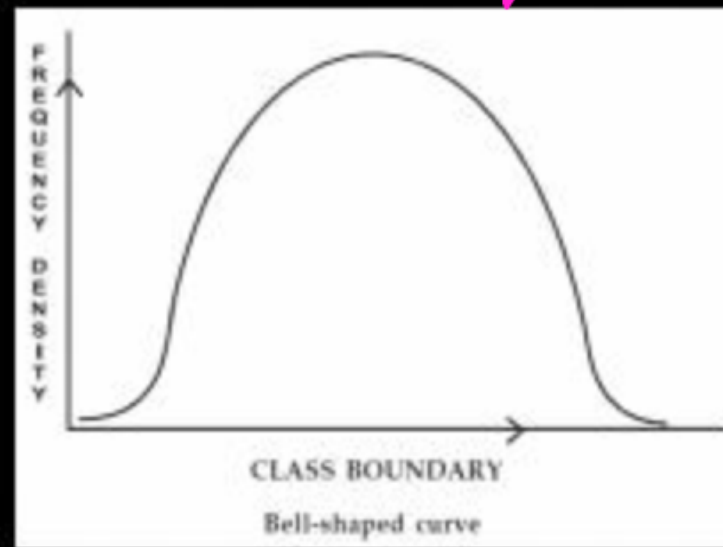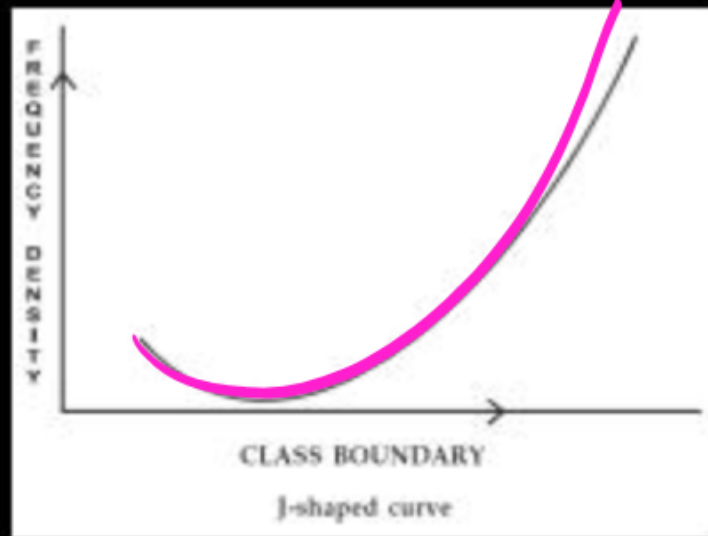More than Ogive
y L.T
C.F

No. of workers / Weekly wages / Median

Frequency Curve

$$F.D = \frac{Freq}{C.L}$$

- Limiting form of frequency polygon or histogram.

- Total area of frequency curve is taken to be unity.

- Graph is between frequency density on vertical axis with class boundary on horizontal axis

Area of freq = 1

J-shaped curve

Bell-shaped curve

U-shaped curve

Mixed Curve
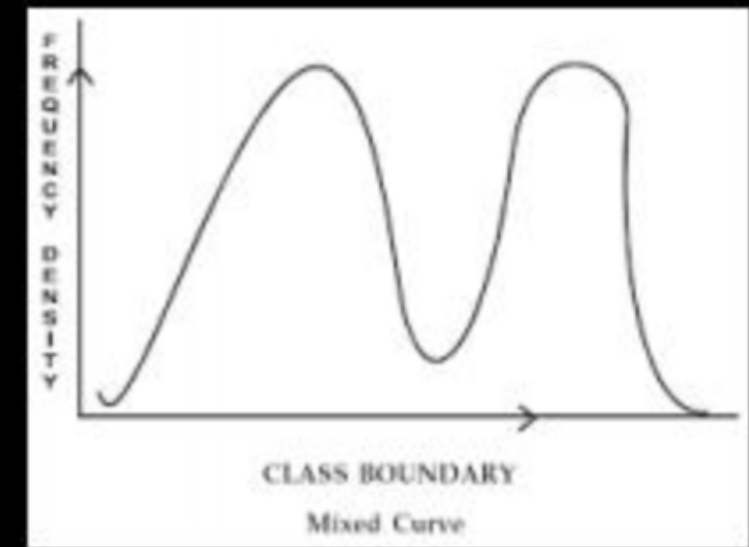
Class	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency	5	8	15	6	4

For the class 20-30, cumulative frequency is

A) 20

B) 13

C) 15

D) 28

Handwritten work:

(Less than) C.f / More than C.f

Class	freq	C.f	C.f
0-10	5	5	38
10-20	8	13	33
20-30	15	28	25
30-40	6	34	10
40-50	4	38	4

5+33=38

Total freq = 38

Mutually exclusive classification

Mutually exclusive :- Cont. Series
↓
C.B and C.L are same

A Excludes both the class limits

B Excludes the upper class limit but includes the lower class limit

C Includes the upper class limit but excludes the upper class limit

D Either (b) or (c)

The LCB is

A An upper limit to LCL

B A lower limit to LCL

C (a) and (b)

D (a) or (b)

$$L.C.B = L.C.L - \frac{gap}{2}$$

$$U.C.B = U.C.L + \frac{gap}{2}$$

Length of a class is

A The difference between the UCB and LCB of that class

B The difference between the UCL and LCL of that class

C (a) or (b)

D Both (a) and (b)

For a particular class boundary, the less than cumulative frequency and more than cumulative frequency add up to

$$L \cdot T \cdot C \cdot F + M \cdot T \cdot C \cdot F = T F$$

A Total frequency

B Fifty per cent of the total frequency

C (a) or (b)

D None of these

Frequency density corresponding to a class interval is the ratio of

$\Rightarrow \dfrac{Freq.}{Class\ Length}$

A Class frequency to the total frequency ✗

B Class frequency to the class length ✓

C Class length to the class frequency ✗

D Class frequency to the cumulative frequency ✗

Relative frequency for a particular class

A Lies between 0 and 1

B Lies between 0 and 1, both inclusive

C Lies between –1 and 0

D Lies between –1 to 1

$$R.F = \frac{Freq\ of\ Class}{Total\ freq}$$

$$\frac{5}{10}$$

Frequency density corresponding to a class interval is the ratio of

A Class frequency to the total frequency

B Class frequency to the class length

C Class length to the class frequency

D Class frequency to the cumulative frequency

$$FD = \frac{Freq}{Class\ length}$$

Mode of a distribution can be obtained from

A Histogram

B Less than type ogives

C More than type ogives

D Frequency polygon

Median of a distribution can be obtained from

A Frequency polygon

B Histogram

C Less than type ogives

D None of these
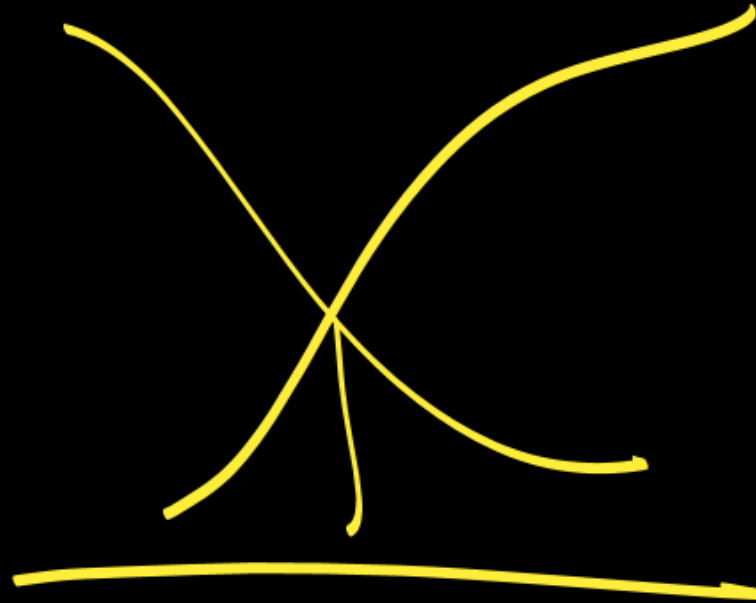
The distribution of profits of a company follows

A J–shaped frequency curve

B U–shaped frequency curve

C Bell–shaped frequency curve

D Any of these

What is Central Tendency

Measure of Central Tendency

→ A M (Commonly → Average)
→ G M
→ H M

→ Median / Partion Value

→ Mode

5ft 5·3ft 5·6ft 5·9ft 6ft 6·3ft - -

x (height)
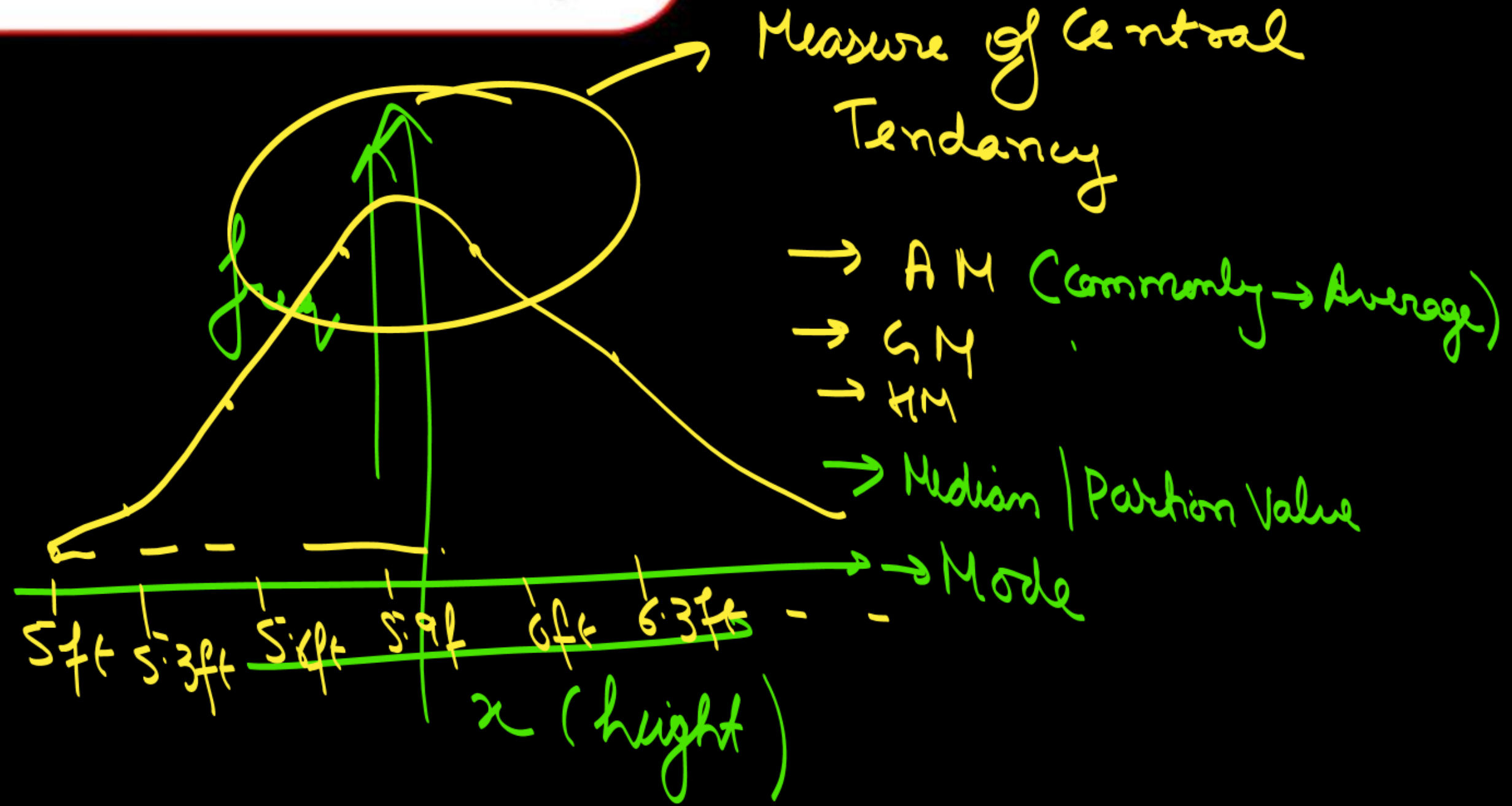
Arithmetic Mean

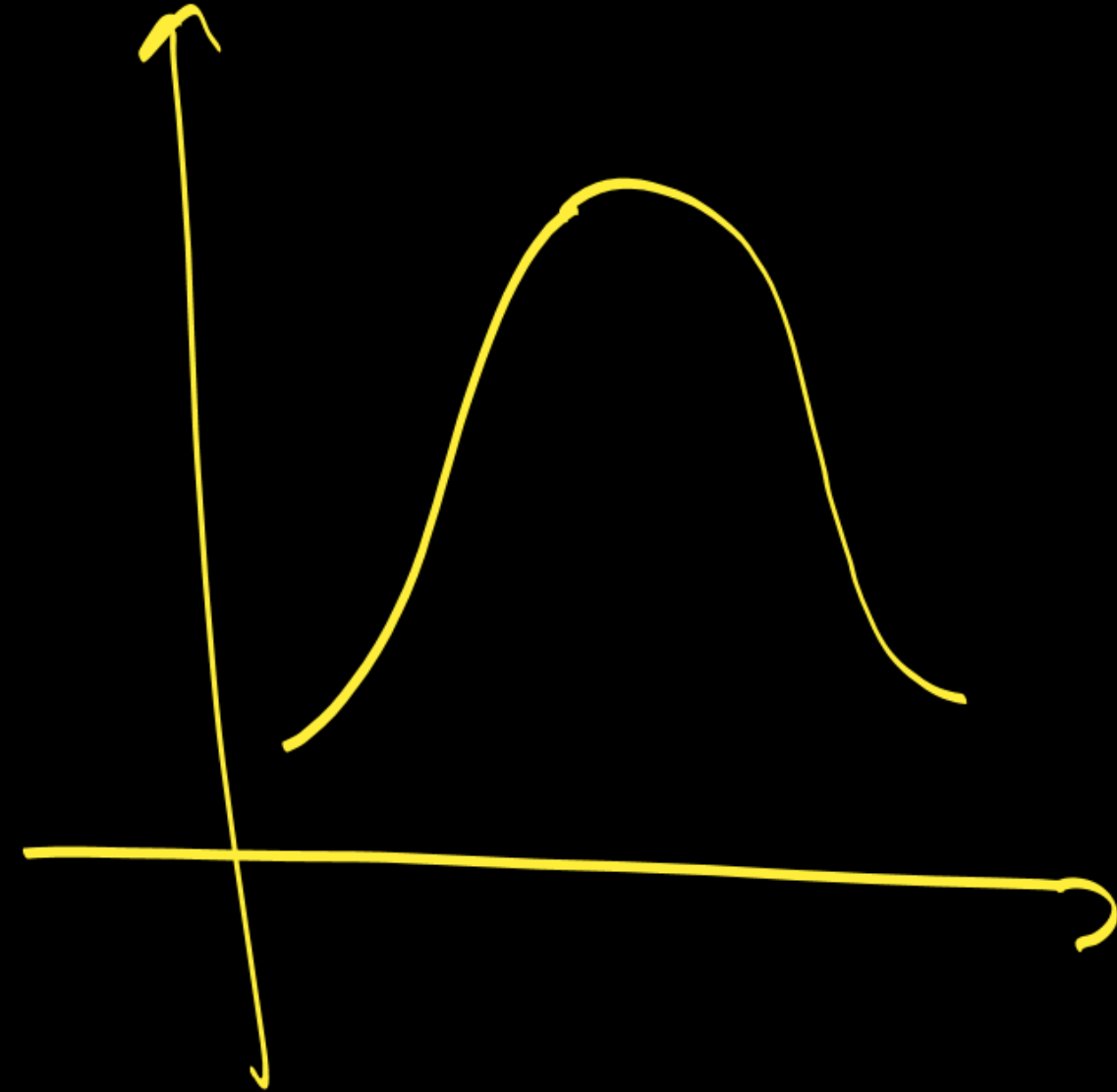- Arithmetic mean is commonly known as average.

- The average of a given set of numbers is called the arithmetic mean, or simply, the mean of the given numbers.

  i.e., Arithmetic Mean (Mean) = $\dfrac{x_1 + x_2 + x_3 + x_4 \cdots \cdots + x_n}{n}$

- Let's variable x has n values- $x_1, x_2, x_3, x_4, x_5, \ldots\ldots x_n$

- Then, Arithmetic Mean is denoted by, which is equal to : $\bar{x} = \dfrac{\sum\limits_{i=1}^{n} x_i}{n}$ or $\dfrac{\sum x_i}{n}$

The following figures give the marks of 10 students in a class test: Marks obtained:
12 8 17 13 15 9 18 11 6 1. Find the arithmetic mean.

$$\bar{x} = \frac{12 + 8 + 17 + 13 + 15 + 9 + 18 + 11 + 6 + 1}{10}$$

$$\bar{x} = \frac{110}{10} = 11$$

Arithmetic Mean

$$k, \quad k, \quad k, \quad k - - \quad \Rightarrow \bar{x} = k$$

- If all the observations assumed by a variable are constants, say k, then the

  AM is _k_

- If all the observations are added, subtracted or multiplied by value k, then

  AM also gets added, subtracted or multiplied by value _k_

- The algebraic sum of deviations of a set of observations from their AM is zero

- AM is affected due to change in origin and/or scale i.e., y = a + bx, then the AM of y
  is given by $\bar{y} =$

$$y = \boxed{a} + b \cdot x \quad \longrightarrow \text{change in origin}$$
$$\longrightarrow \text{change in scale}$$

$$\bar{y} = a + b \cdot \bar{x}$$

- A.M. can not be represented graphically.

- $\Sigma(x - \bar{x})^2 = \text{minimum}$

Sum of deviation from A.M. = 0

$$\Sigma(x_i - \bar{x}) = 0$$

x_i	$x_i - \bar{x}$
1.	$1 - 3 = -2$
2	$2 - 3 = -1$
3	$3 - 3 = 0$
4	$4 - 3 = 1$
5	$5 - 3 = 2$

$$\bar{x} = \frac{15}{5} = 3$$

$$\Sigma(x_i - \bar{x}) = -2 + (-1) + 0 + 1 + (2)$$
$$= -2 - 1 + 0 + 1 + 2 = 0.$$

$$\bar{x} = \frac{2 + 2 + 2 + 2 + 2}{5} = \frac{10}{5} = 2$$

O

$$y = x \pm k$$
$$\bar{y} = \bar{x} \pm k$$

$$y = x \cdot k$$
$$\bar{y} = \bar{x} \cdot k$$

$$y = \frac{x}{k}$$
$$\bar{y} = \frac{\bar{x}}{k}$$

x y

$1 + 2$
$2 + 2$
$3 + 2$
$4 + 2$
$5 + 2$

$$\bar{x} = \frac{15}{3}$$

$$\bar{y} = \frac{25}{5} = 5$$

The arithmetic mean of a set of 5 observations 5, 10, 15, 20 and 25 is 15 However, if each item is increased by 3, then arithmetic mean will be?

$+3 \quad +3 \quad +3 \quad +3 \quad +3$

$\bar{x} = 15$

$\bar{y} = \bar{x} + 3 = 15 + 3 = 18$

The arithmetic mean of a set of 5 observations: 2, 4, 6, 8, 10 is 6. If each item is multiplied by 2, then new arithmetic mean will be?

$\times 2 \quad \times 2 \quad \times 2 \quad \times 2 \quad \times 2$

$\bar{x} = 6$

$\bar{y} = 6 \times 2 = 12$

If it is known that 2 variable x and y are related by equation $2x + 3y = 5$ and $\bar{x} = 1$ then y is?

$$2\bar{x} + 3y = 5$$

$$\bar{x} = 1 \Rightarrow 2\bar{x} + 3\bar{y} = 5$$

$$\Rightarrow 2(1) + 3\bar{y} = 5$$

$$\Rightarrow 2 + 3\bar{y} = 5$$

$$\Rightarrow 3\bar{y} = 5 - 2 = 3$$

$$\Rightarrow \bar{y} = \frac{3}{3} = 1$$

Merits. Arithmetic mean possesses the following merits:

1. It is rigidly defined. ✓

2. It is easy to calculate and simple to understand.

3. It is based on all the observations. ✓

4. It is suitable for further mathematical treatment.

5. Of all the averages, arithmetic mean is affected least by fluctuations of sampling.

$$5.1 \leftarrow 5 \rightarrow 4.8$$

Combined mean

$$\begin{matrix} n_1 & n_2 \\ \bar{x}_1 & \frac{1}{\bar{x}_2} \end{matrix} \Rightarrow \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Correcting the incorrect mean

error

Demerits

$2 \qquad 3 \quad 4 \quad 5 \qquad \boxed{5000}$

$\bar{x} = 3.5$

$$\bar{y} = \frac{2+3+4+5+5000}{5}$$

$= 1002.8$

1. It is very much affected by extreme values.

2. In a distribution with open-end classes the → A.M nhi nikalta h

3. It can neither be determined by inspection nor can it be located graphically.

4. It cannot be computed for a qualitative data such as honesty, beauty, intelligence etc.

↳ Median

5. It may lead to wrong conclusions if the details of the data from which it is obtained are not available

Measures of central tendency for a given set of observations measures

फैला हुआ

A "The scatterness of the observations" → MOD

B "The central location of the observations" → MOCT

C "Both (A) and (B)"

D None of these

While computing the AM from a grouped frequency distribution, we assume that

A The classes are of equal length

B The classes have equal frequency

C All the values of a class are equal to the mid-value of that class

D None of these

$$
\begin{array}{c|c}
 & x_i \\
\hline
0-10 & 5 \\
\hline
10-20 & 15 \\
\hline
20-30 & 25 \\
\hline
30-40 & 35 \\
\end{array}
$$

Which of the following statements is wrong?

A. Mean is rigidly defined

B. Mean is not affected due to sampling fluctuations

C. Mean has some mathematical properties

D. All these

If there are 3 observations 15, 20, 25 then the sum of deviation of the observations from their AM is

$$\sum (x_i - AM) = 0$$

(A) 0

(B) 5

(C) −5

(D) None of these

Median

ascending order
↓
Mid-Value

- Median, for a given set of observations, when arranged in an ascending order or a descending order of magnitude. It may be defined as the middle-most value

- As distinct from the arithmetic mean, which is based on all the items of the distribution, the median is what is called a positional average.

For ungrouped data consisting of n observations, the calculation of median involves the following steps:

a. When n is odd, the value is given by $\left(\dfrac{n+1}{2}\right)^{th}$ term

b. When n is even, the value is given by- $\dfrac{\left(\dfrac{n}{2}\right)^{th} + \left(\dfrac{n}{2}+1\right)^{th}}{2}$

Find the median for the following data:

(i) 19 22 17 20 12

(ii) 58 49 64 70 91 34

(1) $\boxed{19}$ $\boxed{22}$ $\boxed{17}$ $\boxed{20}$ $\boxed{12}$

$\Rightarrow$ 12, 17 $\boxed{19}$, 20, 22

median =

Median
for odd no. of
terms

Mid value = 19

$= \left(\dfrac{n+1}{2}\right)^{th}$ term $= \left(\dfrac{5+1}{2}\right)^{th} = 3^{rd} = 19$

② 34, 49, 58, 64, 70, 91

Median =
no. of terms
is even
(n)

$= \dfrac{\left(\dfrac{n}{2}\right)^{th} + \left(\dfrac{n}{2}+1\right)^{th}}{2}$

$= \dfrac{\left(\dfrac{6}{2}\right)^{th} + \left(\dfrac{6}{2}+1\right)^{th}}{2}$

$= \dfrac{3^{rd} + 4^{th}}{2}$

$= \dfrac{58+64}{2} = \dfrac{122}{2} = 61$

In case of an even number of observations which of the following is median?

A Any of the two middle – most value

B The simple average of these two middle values $= \dfrac{\left(\dfrac{n}{2}\right)^{th} + \left(\dfrac{n}{2}+1\right)^{th}}{2}$

C The weighted average of these two middle values

D Any of these

(i) If x and y are two variables, to be related by y = a + bx for any two constants a and b, then the median of y is given by

$$y = a + bx$$

$$\Rightarrow y_{med} = a + bx_{med}$$

(ii) For a set of observations, the sum of absolute deviations is minimum when the deviations are taken from the median.
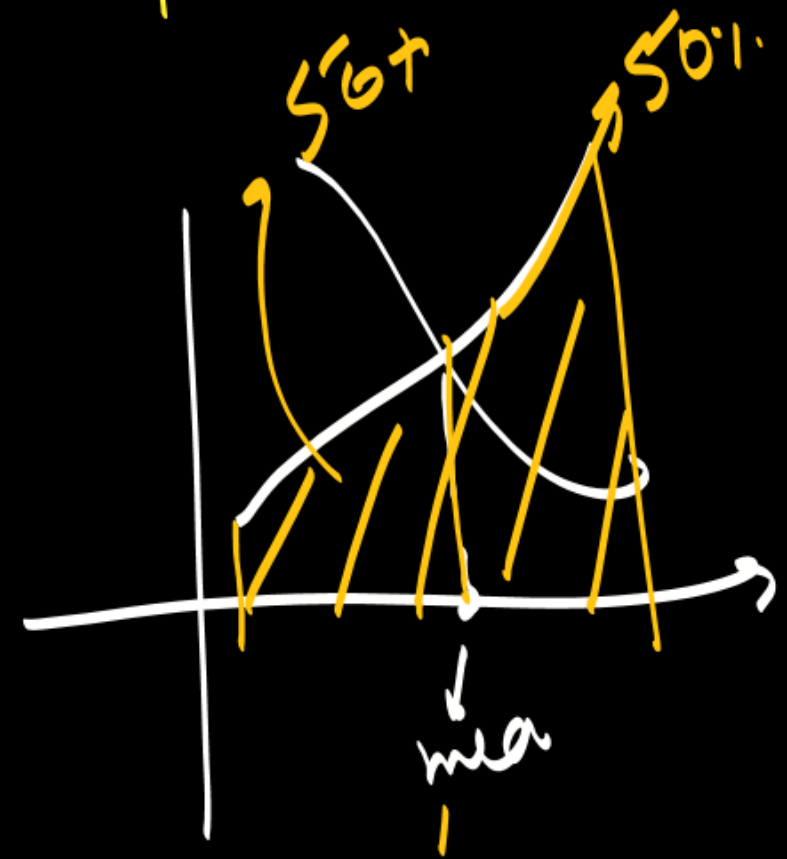
$$\Sigma |x - x_{med}| \Rightarrow minimum$$

$$19 \quad |19-19| = |0| = 0$$

$$22 \quad |22-19| = |3| = 3$$

$$20 \quad |20-19| = |1| = 1$$

$$17 \quad |17-19| = |-2| = 2$$

$$12 \quad |12-19| = |-7| = 7$$

$$\Sigma |x_i - x_{med}|$$

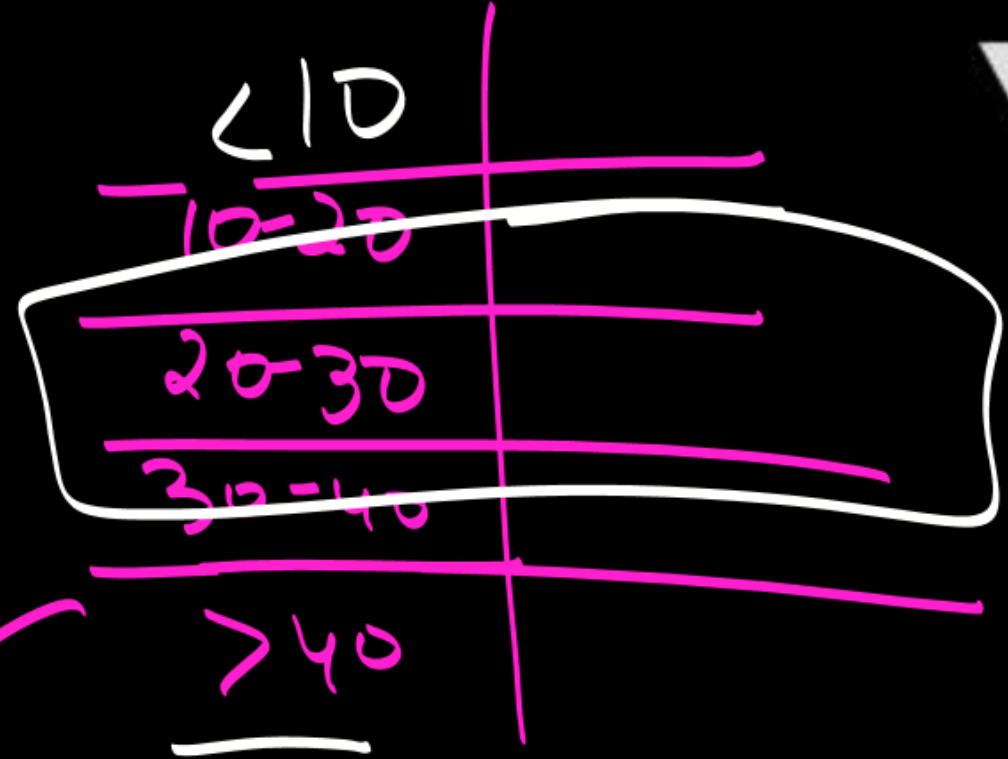$$\Sigma |x_i - x_{med}| = 13 \rightarrow \text{min pos}$$

$$x_{med} = 19$$

Merits of Median

1. It is rigidly defined. ✓

2. It is easy to calculate and simple to understand. ✓

3. It can be computed while dealing with a distribution with open end classes

4. Being a- *centre position* positional average, it is not much affected by extreme observations.

5. It is the most appropriate average to be used while dealing with qualitative data.

6. It can sometimes be located by inspection and can also be determined graphically.

< 10
$10 - 20$
$20 - 30$
$30 - 40$
> 40

2   5   ⑦   10   15000

Ogives

Demerits of Median

(Handwritten at top: 25% | 25% | 25% | 25%)

(Handwritten: Median)

1. Median, being a positional average, is not based on each and every item of the distribution.

   (Handwritten: → Median only 50% of total data)

2. It is not suitable for further mathematical treatment.

3. It cannot be determined exactly for an ungrouped data consisting of an even number of observations.

4. In comparison to arithmetic mean, it is much affected by sampling fluctuations.

5. For calculating median, it is necessary to arrange the data in order of magnitude.

Two variables x and y are given by $y = 2x - 3$. If the median of x is 20, what is the median of y?

(A) 20

(B) 40

(C) 37

(D) 35

$y = 2x - 3$

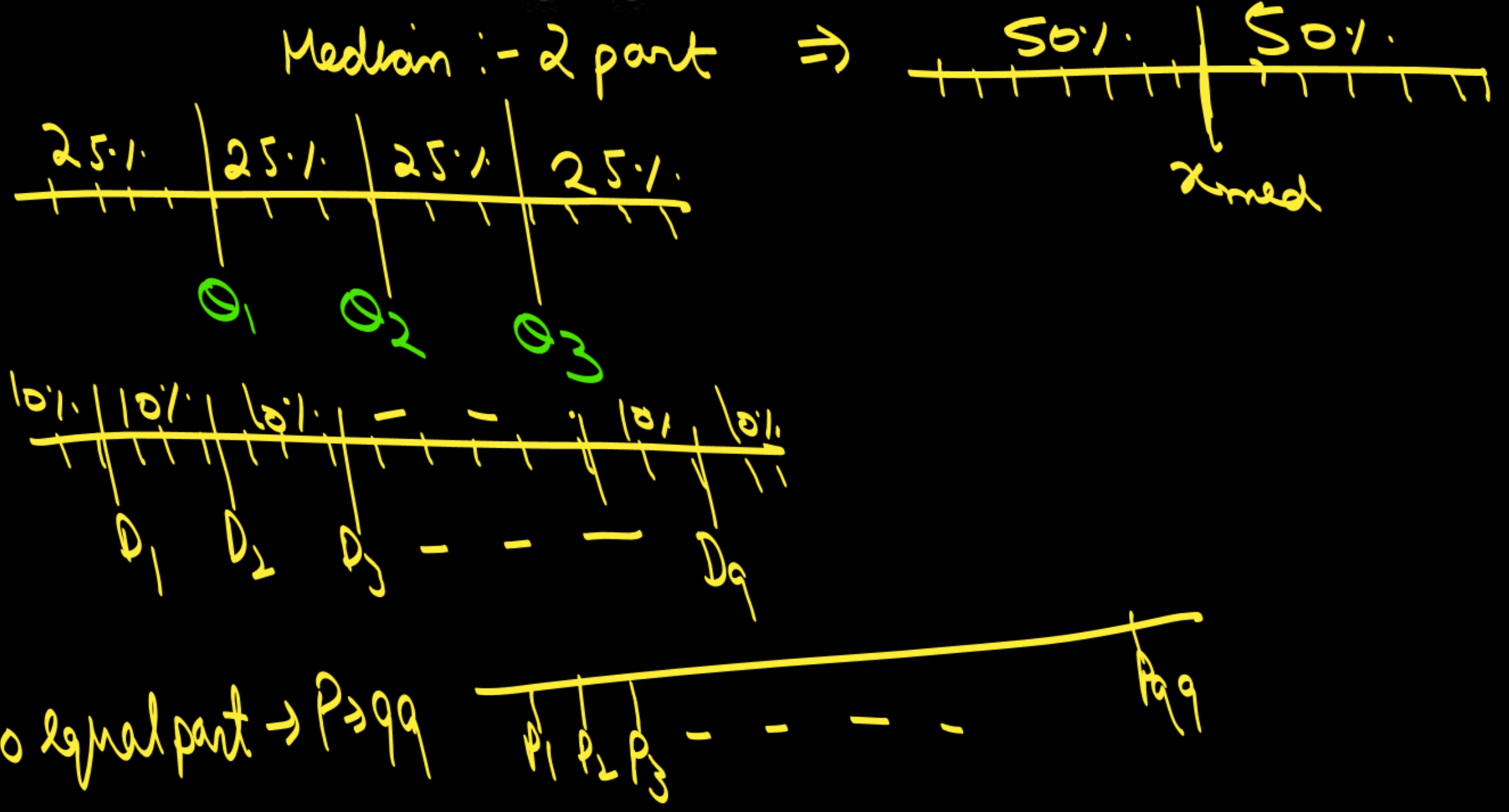$x_{med} = 20$

$y_{med} = 2 x_{med} - 3$

$= 2(20) - 3 = 40 - 3 = 37$

Partition Value or Quartiles or Deciles or Percentiles

These may be defined as values dividing a given set of observations into a number of equal parts.

Median :- 2 part $\Rightarrow$

50% | 50%

x_{med}

- **Quartiles** –

4 equal part $\Big\}$ 3Q.

25% | 25% | 25% | 25%

Q_1 Q_2 Q_3

- **Deciles** –

10 equal part $\Big\}$ 9D.

10% | 10% | 10% | --- | 10% | 10%

D_1 D_2 D_3 --- --- --- D_9

- **Percentiles** -

100 equal part $\rightarrow$ P$\geq$99

P_1 P_2 P_3 --- --- --- P_{99}

- For ungrouped data consisting of n observations (not necessarily all distinct), the calculation of kth quartile $Q_k = (k = 1, 2, 3,)$ involves the following steps:

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} \text{term}$$

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} \text{term}$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \text{term}$$

$$Q_k = \left(\frac{k(n+1)}{4}\right)^{th}$$

Calculate the quartiles Q_1, Q_2, Q_3 for the following data: -

23, 11, 9, 8, 6

Ascending order :- 6, 8, (9), 11, 23

$n = 5$

$$Q_1 = \left(\frac{1 \times (5+1)}{4}\right)^{th} = 1.5^{th} = 1^{st} + 0.5(2^{nd} - 1^{st}) = 6 + 0.5(8-6)$$
$$= 6 + 0.5(2) = 7$$

Median $= Q_2 = \left(\frac{2 \times (5+1)}{4}\right)^{th} = 3^{rd} = 9$

$$Q_3 = \left(\frac{3(5+1)}{4}\right)^{th} = 4.5^{th} = 4^{th} + 0.5(5^{th} - 4^{th})$$
$$= 11 + 0.5(23-11)$$
$$= 11 + 6 = \underline{17}$$

Quartiles are the values dividing a given set of observations into

A Two equal parts

B Four equal parts $\rightarrow Q_1, Q_2, Q_3$

C Five equal parts

D None of these
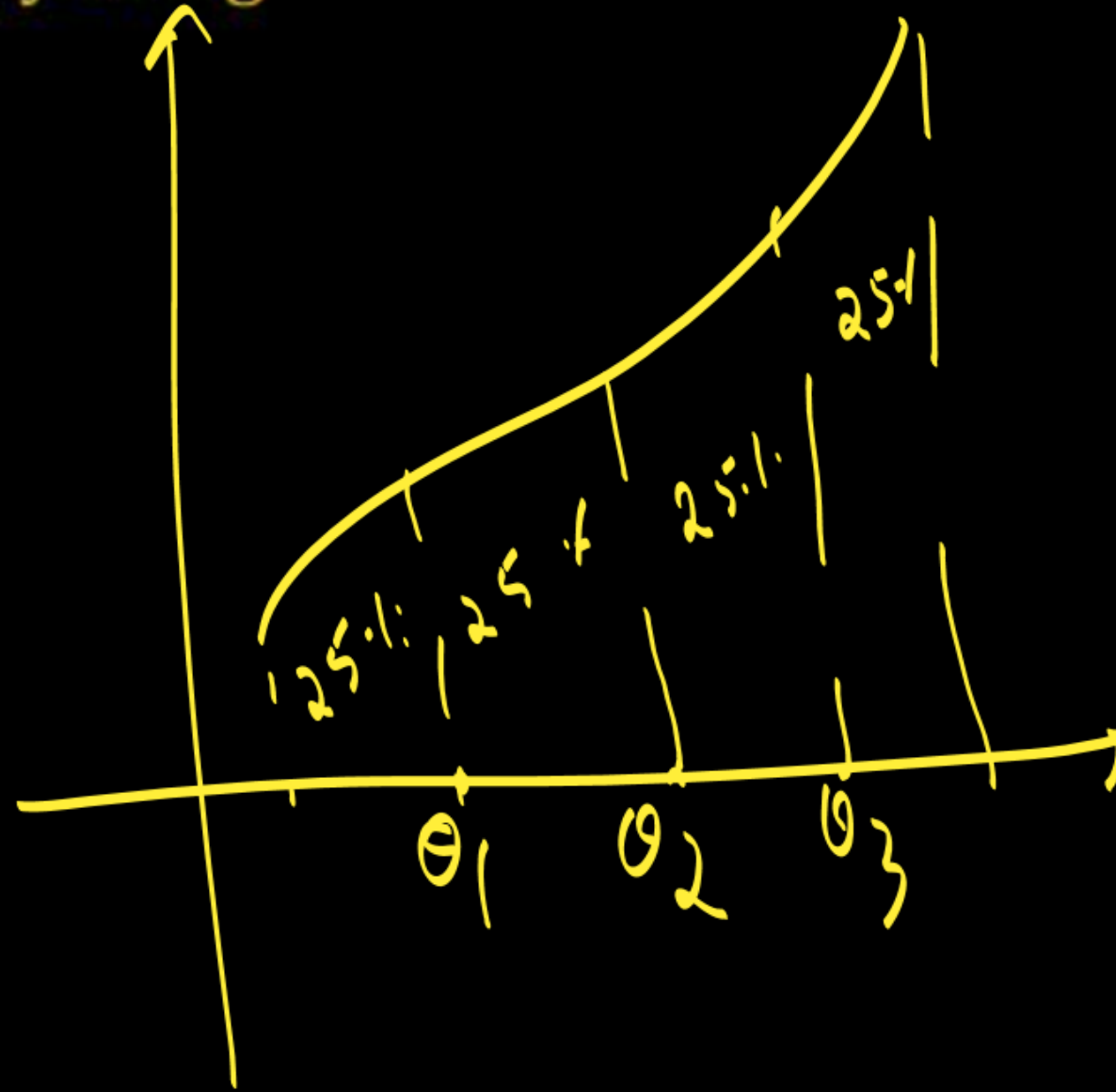
Decile

$$D_K = \left(\frac{K(n+1)}{10} \right)^{th}$$

Percentile

$$P_K = \left(\frac{K(n+1)}{100} \right)^{th}$$
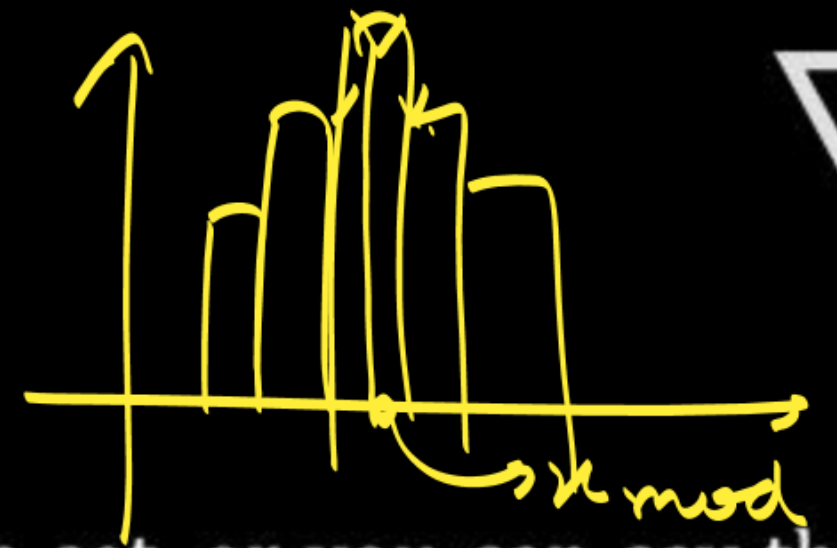
Quartiles can be determined graphically using

A Histogram

B Frequency Polygon

C Ogive

D Pie chart

▶ Mode

- The mode is the value that occurs the most often in a data set, or you can say the data with the maximum frequency is called as mode.

EXAMPLE :

1, 2, 3, 4, 4, 5, 2, 3, 4 What will be mode?

4 → no. of times ⇒ 3

Mode = 4

- **Bi- Modal Distribution**   **Multi- Modal Distribution**   **No Mode**

1,1, 2, 3, 4, 4, 4, 3,3   1, 2, 3, 4, 5

- It is not uniquely defined.

- Calculated using Histogram ✓✓

Properties of Mode

> We also note that if $y = a + bx$, then $y_{mo} = a + bx_{mo}$

$$y = a + bx$$

$$y_{mod} = a + bx_{mod}$$

If $y = 2 + 1.50x$ and mode of x is 15, what is the mode of y?

$$x_{mod} = 15$$

$$y_{mod} = 2 + 1.5 x_{mod} = 2 + 1.5(15) = 2 + 22.5 = 24.5$$

Merits and demerits of mode

Merits. Mode possesses the following merits

1. It is simple to understand and easy to calculate.

2. In some cases it can be located merely by inspection.

3. It can be determined graphically from a histogram.

4. It is not at all affected by extreme observations and can be calculated even if extreme values are not known.

5. It can be conveniently determined for distribution with open end classes.

x_i	f_i
2	5
4	10
5	8
10	2
1,0000	1

Demerits of Mode

Mode has the following drawbacks

1. It is not rigidly defined → no single value

2. It is not based on all the observations.

3. It is not suitable for further mathematical treatment.

4. As compared to mean, mode is affected to a greater extent by the fluctuations of sampling.

5. The value of mode cannot always be determined. In some cases, we may have a bi-modal distribution.

Which of the following measure(s) satisfies (satisfy) a linear relationship between two variables?

$$y = a + bx$$

(A) Mean

(B) Median

(C) Mode

(D) All of these
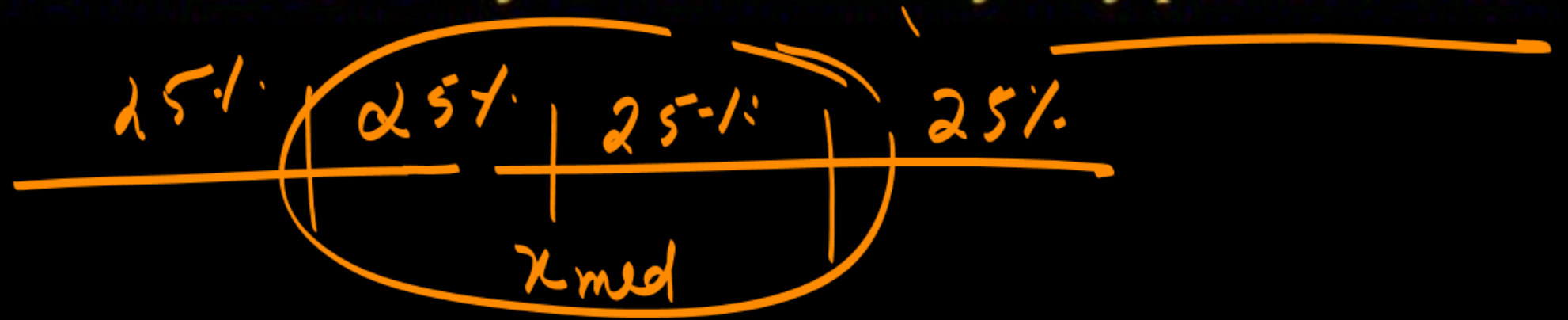
Which of the following measures of central tendency is based on only fifty percent of the central values?

A Mean

B Median

C Mode

D Both (A) and (B)

25% 25% 25% 25%

x_{med}

Geometric Mean

$$GM = (x_1 \times x_2 \times x_3 - - \times x_n)^{\frac{1}{n}}$$

- GM is best measure of CT for ratio & percentages are constant. For example, with such data as rates of change, per cent increase in sales, population sizes over consecutive time periods etc.

- G.M. of two numbers 4 and 9 is

$$GM = (4 \times 9)^{\frac{1}{2}} = (36)^{\frac{1}{2}} = 6$$

- And, G.M. of three numbers 1, 4 and 128 is

$$\Rightarrow GM = (1 \times 4 \times 128)^{\frac{1}{3}} = (512)^{\frac{1}{3}} = 8$$

$$5|2$$
$$\sqrt{} \rightarrow 12 \text{ times}$$
$$\frac{1}{n}$$
$$\frac{1}{5} x = \frac{1}{12} \text{ times}$$

- It is the most appropriate average to be used in the construction of index numbers.

- It is the most suitable average to be used when it is desired to give more weightage to smaller items and vice-versa.

Properties

$$GM = (2 \times 2 \times 2 \times 2)^{\frac{1}{4}}$$
$$= (2^4)^{\frac{1}{4}} = 2$$

1. If all the observations assumed by a variable are constants, say $K > 0$, then the GM of the observations is also K.

$$\to K, K, K - - \Rightarrow GM = K$$

2. GM of the product of two variables is the product of their GM's i.e. if $z = xy$, GM of z =

$$z = x \times y$$
$$GM_z = GM_x \times GM_y$$

$$z = xy \to GM_z = GM_x \cdot GM_y$$

$$z = \frac{x}{y} \Rightarrow GM_z = \frac{GM_x}{GM_y}$$

3. GM of the ratio of two variables is the ratio of the GM's of the two variables i.e. if $z = x/y$

GM of z=

Which of the following measure of the central tendency is difficult to compute?

A Mean

B Median

C Mode

D GM

Harmonic Mean

$$HM = \frac{\text{total obs.}}{\text{Sum of reciprocal of obs.}}$$

- If all the observations taken by a variable are constants, say k, then the HM of the observations is also_____

- If there are two groups with n_1 and n_2 observations and H_1 and H_2 as respective HM's than the combined HM is given by

  n_1 n_2

  HM_1 HM_2

  $$HM = \frac{n_1 + n_2}{\dfrac{n_1}{HM_1} + \dfrac{n_2}{HM_2}}$$

- Harmonic means is used for reciprocal relationship average of speed, finding P/E ratio.

Find the harmonic mean of 5 numbers 4, 5, 6, 10 and 12.

$$n = 5$$

$$HM = \dfrac{5}{\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{10} + \frac{1}{12}} = \dfrac{5}{x} = 5 \times \frac{1}{x} = 6.25$$

Cal:-

AC

$$1 \div 4 =$$
$$1 \div 5 =$$
$$1 \div 6 =$$
$$1 \div 10 =$$
$$1 \div 12 =$$

$$x = 0.8$$

or $\dfrac{1}{x} \Rightarrow \boxed{0.8 \div = } \times 5 \Rightarrow Any$

$$\times 5$$

Properties

1. If all the observations taken by a variable are constants, say k, then the HM of the observations is also k.

$$HM = \frac{4}{\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}} = \frac{4}{2} = 2$$

$2 \qquad 2 \qquad 2 \qquad 2$

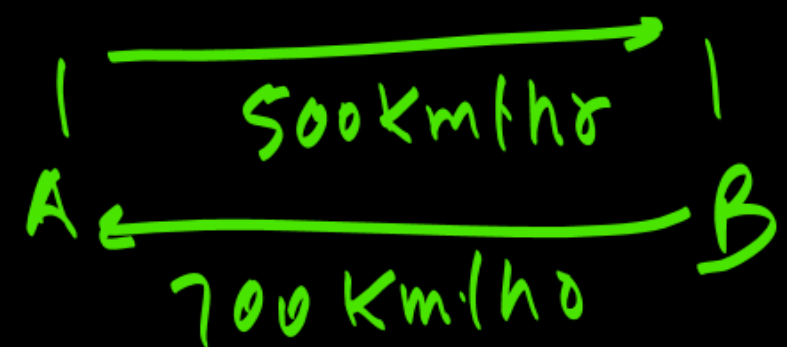2. If there are two groups with n_1 and n_2 observations and H_1 and H_2 as respective HM's than the combined HM is given by

An aeroplane flies from A to B at the rate of 500 km/hour and comes back from B to A at the rate of 700 km/hour. The average speed of the aeroplane is

A 100 km per hour

B 583.33 km. per hour

C $100 \sqrt{35}$ km. per hour

D 620 km. per hour.

$$A \xleftarrow{\quad 500 \, km/hr \quad} B$$
$$700 \, km \cdot l h \, \sigma$$

$$HM = \frac{2}{\frac{1}{500} + \frac{1}{700}} = \frac{2}{x} = \frac{2 \times 1}{x} = 583.33 \, km/hr$$

Calc:-
$1 \div 500 =$
$1 \div 700 =$
$2 \times \frac{1}{x} = (x \div =) \times 2 \Rightarrow Ans$

Which of the following measure(s) possesses (possess) mathematical properties?

A AM

B GM

C HM

D All of these

Important Pointers

- Positional Average is $\boxed{\text{Median and Mode}}$

- Quantitative Average is AM HM and GM

- $\text{AM} \geq \text{GM} \geq \text{HM}$

  $\rightarrow$ AM > GM > HM $\Rightarrow$ when values are distinct

  AM = GM = HM $\Rightarrow$ If all are same

- $\text{GM} = \sqrt{\text{AM} \times \text{GM}}$

  Moderately Skewed Dist

  Mean - Mode = 3 (Mean - Median)

Empirical Relationship :

- Mode = 3Median - 2*Mean*

  or Mode –*Mean*= 3(Mode - *Mean*)

What is Central Tendency

	ARITHMETIC MEAN	GEOMETRIC MEAN	HARMONIC MEAN	MODE
Individual Observation	$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$ $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$	$GM = (x_1 \times x_2 \times x_3 \ldots \times x_n)^{1/n}$ Logarithm of G for a set of observations is the AM of the logarithm of the observations; i.e. $$\log GM = \frac{\sum \log x}{n}$$ $$G.M. = \text{Antilog} \frac{\sum \log x}{n}$$	$$H.M. = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$$	The value that occurs the maximum number of times

What is Central Tendency

	ARITHMETIC MEAN	GEOMETRIC MEAN	HARMONIC MEAN	MODE
Frequency Distribution	$\bar{x} = \dfrac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{f_1 + f_2 + f_3 + \cdots + f_n}$	$GM =$ $(x_1 f_1 \times x_2 f_2 \times x_3 f_3 \ldots \times x_n f_n)^{1/n}$	$H.M. = \dfrac{N}{\Sigma\left(\frac{f_i}{x_i}\right)}$	Mode $= l + \left(\dfrac{f_0 - f_1}{2f_0 - f_1 - f_{-1}}\right) \times C$ Where, l_1 = LCB of the modal class i.e. the class containing mode. f_0 = frequency of the modal class f_{-1} = frequency of the pre-modal class f_1 = frequency of the post modal class C = class length of the modal class

Central Tendency

	ARITHMETIC MEAN	GEOMETRIC MEAN	HARMONIC MEAN	MODE

Relationship variables	$\bar{y} = a + b\bar{x}$	if $z = xy$, then
		GM of $z = (GM \text{ of } x) \times (GM \text{ of } y)$
		if $z = x/y$ then
		GM of $z = (GM \text{ of } x)/(GM \text{ of } y)$
				$y_{mo} = a + bx_{mo}$

Weighted Mean	Weighted A.M $= \dfrac{\Sigma x_i w_t}{\Sigma w_t}$	Weighted G.M $= \text{Antilog} \dfrac{\Sigma w_i \log x_i}{\Sigma w_i}$	Weighted $H.M = \dfrac{\Sigma w_i}{\Sigma \left(\frac{w_i}{x_i}\right)}$	
Combined Mean	Combined $A \cdot M \bar{x}_{12} = \dfrac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$		Combined $H.M = \dfrac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}$	

Partition Value

	MEDIAN	QUARTILES (Q_1, Q_2 & Q_3)	DECILES (D_1, D_2, D_3 ....D_9)	PERCENTILES (P_1, P_2, P_3,....., P_{99})
Discrete Series	Median = Size of $\left(\frac{N+1}{2}\right)^{th}$ item	Q_1 quartile is given by the $\frac{1}{4}(N+1)$ th value the Q_n quartile is given by the $\frac{n}{4}(N+1)$ th value	The D_1 Decile is given by the $\frac{1}{10}(N+1)$ th value D_n Decile is given by the $\frac{n}{10}(N+1)th$ value	The P_1 Percentile is given by the $\frac{1}{100}(N+1)$ th value P_n Percentile is given by the $\frac{n}{100}(N+1)$ th value
Group Frequency Distribution	Median $= l_1 + \left(\frac{\frac{N}{2}-CF}{f}\right) \times C$ l_1 = lower class boundary of the median class i.e. the class containing median. N = total frequency. CF = less than cumulative frequency corresponding to $l1$. (Pre median class) f= frequency of the median class $C = l_2 - l_1$ = length of the median class. $y_{me} = a + bx_{me}$	$Q_n = l_1 + \left(\frac{N \cdot p - CF_l}{f}\right) \times C$ l^1 = lower class boundaty of the Quartile class i.e. the class containing Quartile. N = total frequency. p = $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ for Q_1, Q_2, Q_3 respectively CF = less than cumulative frequency corresponding to $l1$. (Pre Quartile class) F = frequency of the Quartile class. $C = l_2 - l_1$ = length of the Quartile class.	$D_n = l_1 + \left(\frac{N \cdot p - CF_l}{f}\right) \times C$ l_1 = lower class boundary of the Decile class i.e. the class containing Decile. N = total frequency. $p = \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, ..., \frac{9}{10}$ for D_1, D_2 , $D_3, ..., D_9$ respectively CF = less than cumulative frequency corresponding to l_1. (Pre Decile class) F = frequency of the Decile class. $C = l_2 - l_1$ = length of the Decile class.	$P_n = l_1 + \left(\frac{N \cdot p - CF_l}{f}\right) \times C$ l_1 = lower class boundary of the Percentile class i.e. the class containing Percentile. N = total frequency $p = \frac{1}{100}, \frac{2}{100}, \frac{3}{100}, ..., \frac{99}{100}$ for P_1 $P_2, P_3, ..., P_{99}$ respectively CF = less than cumulative frequency corresponding to l_1. (Pre Percentile class) F = frequency of the Decile class. $C = l_2 - l_1$ = length of the Percentile class.

Note:- 1. $y_{me} = a + bx_{me}$
2. $\sum(x_i - A)$ is minimum if we choose A as the median.

Measure of Dispersion

What is Dispersion?

Dispersion in statistics is a way of describing how spread out a set of data is.

Consider the following three sets of observations, each containing 9 items:

										Total	Mean (AM)
Set A	20	20	20	20	20	20	20	20	20	180	20
Set B	16	17	18	19	20	21	22	23	24	180	20
Set C	12	14	16	18	20	22	24	26	28	180	20

Therefore, we can say that, we need some more measures in addition to central tendency to describe the data completely.
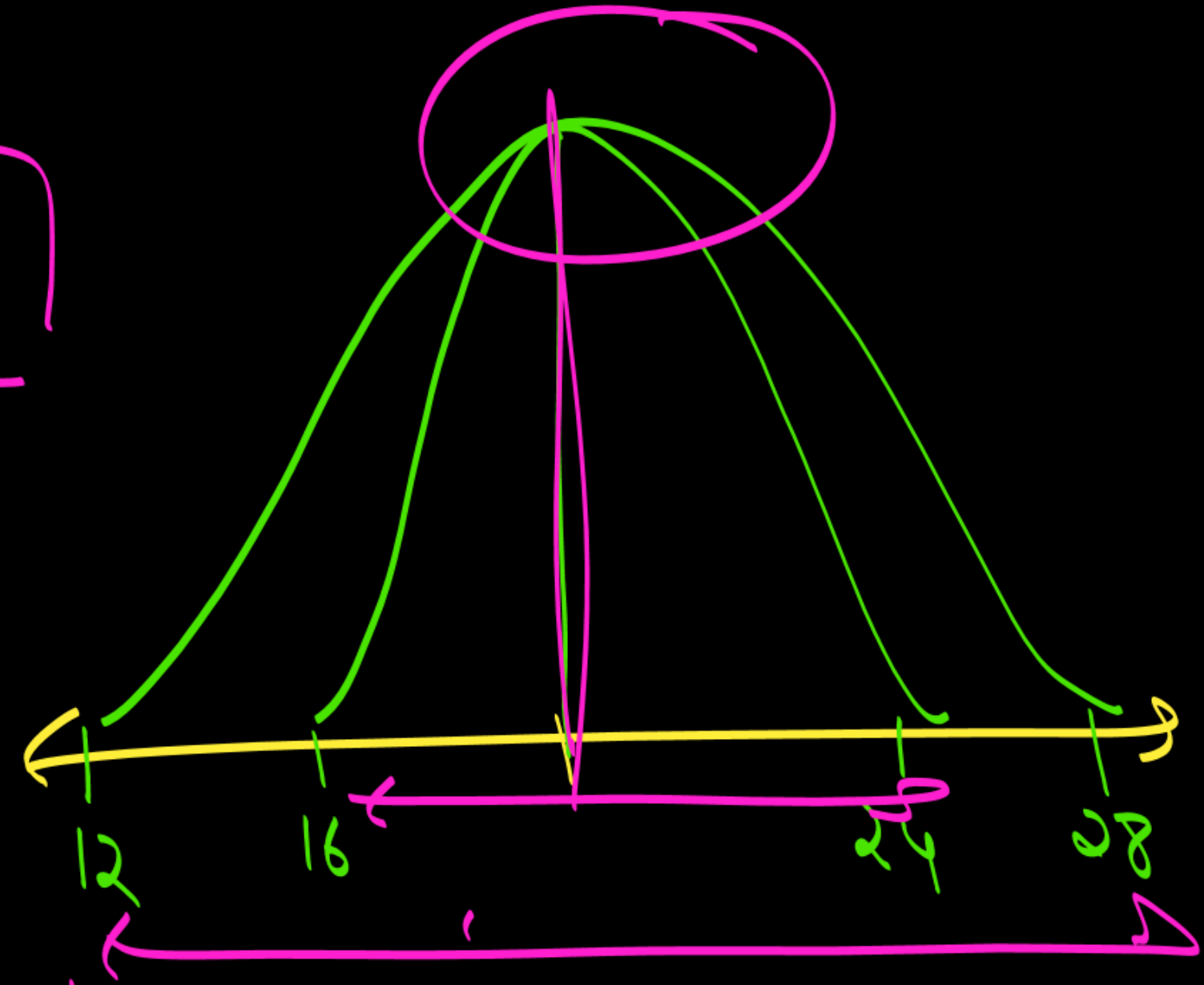
फैलाव

Dispersion

$\downarrow$

Measure of Dispersion

$\rightarrow$ Range

12   16   24   28

MD | QD | SD

Measures of dispersion

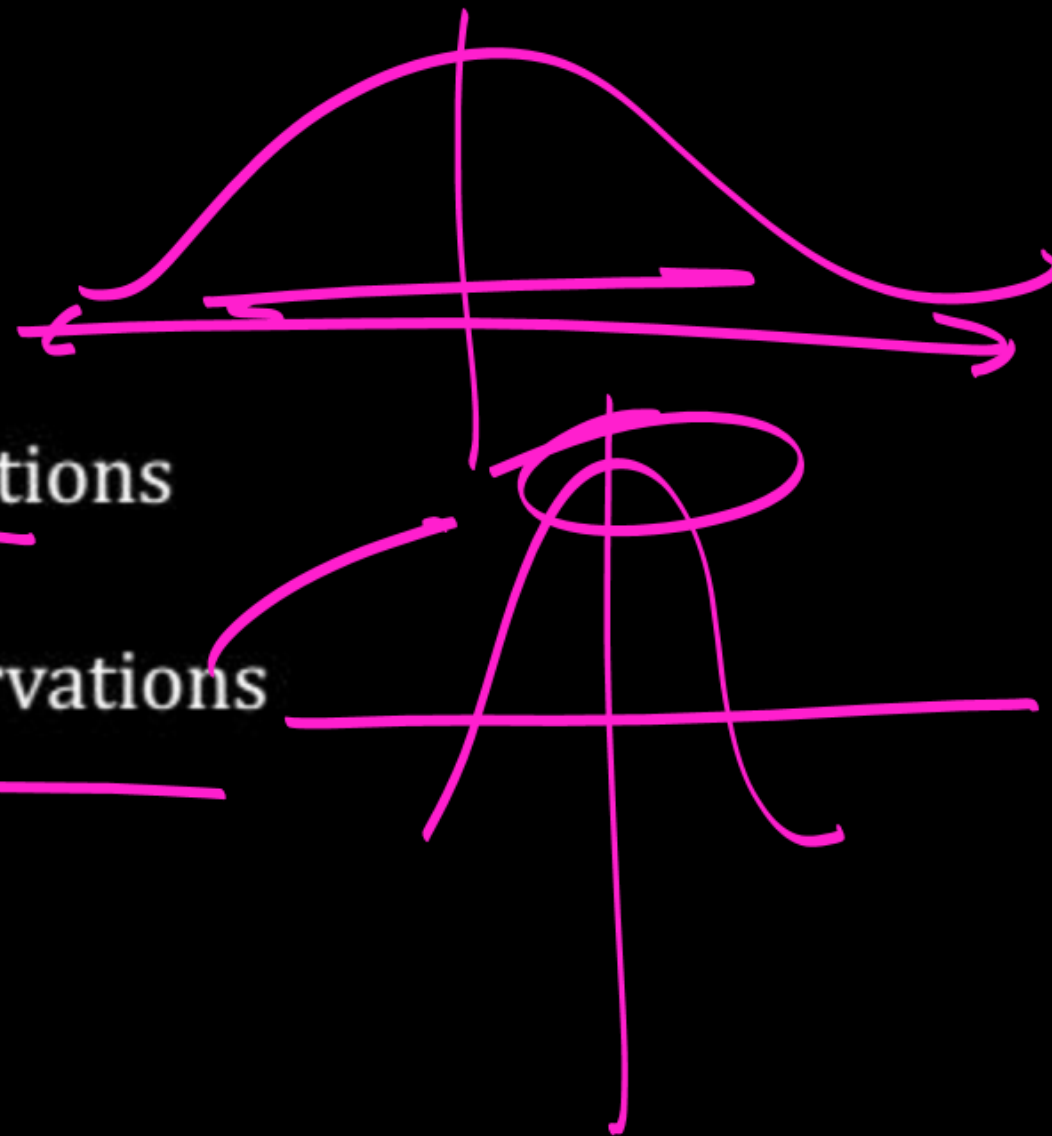| ABSOLUTE MEASURES OF DISPERSION	RELATIVE MEASURES OF DISPERSION
• Absolute measures are dependent on the unit of the variable under consideration	• Relative measures of dispersion are unit free.
• Easy to comprehend and Compute	• For comparing two or more distributions, relative measures of dispersion are considered.
Different Measures ways: -	Different Measures ways: -
• Range	• Coefficient of Range.
• Mean Deviation	• Coefficient of Mean Deviation
• Standard Deviation	• Coefficient of Variation
• Quartile Deviation	• Coefficient of Quartile Deviation.

For comparision of diff relatable data (handwritten annotation)

Dispersion measures

A. The scatterness of a set of observations

B. The concentration of a set of observations

C. Both (A) and (B)

D. Neither (A) and (B)

When it comes to comparing two or more distributions we consider

A Absolute measures of dispersion

B Relative measures of dispersion

C Both (A) and (B)

D Either (A) or (B)

Measures of Dispersion

	Absolute	Relative	If y = a + bx
RANGE (R)	Range = Largest (L) – Smallest (S)	Co efficient of Range $= \dfrac{L-S}{L+S} \times 100$	$R_y = |b| \times R_x$ ✓✓
MEAN DEVIATION (M.D) about A	M. $D_A = \dfrac{1}{n}\Sigma|x - A|$	Co efficient of M.D from A $= \dfrac{M.D\ about\ A}{A} \times 100$	M.D $D_y = |b| \times$ M.D D_x
MEAN DEVIATION (M.D) about A.M ($\bar{x}$)	M.D about Mean $= \dfrac{1}{n}\Sigma|x_i - \bar{X}|$	Co efficient of M.D from A.M $= \dfrac{M.D\ about\ \bar{x}}{\bar{x}} \times 100$	M. $D_y = |b| \times$ M $\cdot D_x$
MEAN DEVIATION (M.D) about Median	M.D about Median $= \dfrac{1}{n}\Sigma|x_i - Median|$	Co efficient of M.D from Median $= \dfrac{M.D\ about\ A}{A} \times 100$	M.D $D_y = |b| \times$ M $\cdot D_x$

Properties of Range

$$y = \frac{y}{\text{Range } y} = |b| \times \text{Range of } X$$

- Range remains unaffected due to a change of origin but affected in the same ratio due to a change in scale i.e., if for any two constants a and b, two variables x and y are related by

$$y = a + bx,$$

$$\text{Range of } X = |b| \text{ Range of } X$$

Merits

The range possesses the following merits:

- It is simple to understand and easy to calculate.

- It requires minimum time to calculate the value of range.

- It is not based on all the observations. It considers only the extreme values

Demerits

1. It is not based on all the observations.

2. Range is a poor measure of Dispersion. It considers only the extreme values

3. It is very much affected by fluctuations of sampling. Its value varies widely from sample to sample

4. It cannot be calculated for grouped frequency distribution with open-end classes.

5. It is not suitable for further mathematical treatment.

If the relationship between x and y is given by 2x + 5y = 10 and the range of x is 5, what would be the range of y?

$$2x + 5y = 10$$

Range of $x = 5$

Range of $y = ?$

$$2x + 5y = 10$$

$\Rightarrow 2 \text{ Range of } x = 5 \text{ Range of } x$

$\Rightarrow 2(5) = 5 \times \text{Range of } y$

$\Rightarrow$ Range of $y = 2$

$$\Rightarrow \frac{5y}{5} = \frac{10 - 2x}{5}$$

$$\Rightarrow y = \frac{10}{5} - \frac{2}{5}x$$

$$\text{Range of } y = \left|\frac{-2}{5}\right| \times \text{Range of } x$$

$$= \frac{2}{5} \times 5 = 2$$

Mean Deviation

$\rightarrow$ Average of absolute deviation around mean or median

$$MD = \frac{\sum |x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum |x_i - x_{med}|}{n}$$

- Mean deviation is computed about mean or median because mode is ill defined

- Mean deviation about median is beneficial because the sum of the deviations of items from median is least

$$\sum |x_i - x_{med}| \Rightarrow \text{minimum}$$

- Mean is more frequently used in computing the average deviation

- Mean deviation remains unchanged due to a change of origin but changes in the same ratio due to a change in scale

If x and y are related as 3x + 4y +7 = 0 and mean deviation of x is 6.40, what is the mean deviation of y?

$$3x + 4y + 7 = 0$$

$MD_x = 6.4$

$MD_y =$

Shortcut

$3x \oplus 4y + 7 = 0$

$3\,MD_x = 4\,MD_y$

$\Rightarrow 3(6.4) = 4\,MD_y$

$\Rightarrow MD_y = \dfrac{3}{4}(6.4) = 4.8$

$4y = \dfrac{-7 - 3x}{4}$

$y = -\dfrac{7}{4} - \dfrac{3}{4}x$

$MD_y = \left| -\dfrac{3}{4} \right| \times MD_x$

$= \dfrac{3}{4} \times 6.4 = 3 \times 1.6 = 4.8$

$y = \cancel{\alpha} + \dfrac{6}{16}x$

$MD_y = |b| \times MD_x$

Merits of Mean Deviation

1. It is easy to understand and simple to calculate.

2. It is based on each and every item of the data.

3. It is rigidly defined.

4. As compared , it is less affected by extreme observations.

Demerits of Mean Deviation

1. The major drawback of mean deviation is that algebraic <u>signs are ignored while</u> taking the deviations of the items.

2. It is not suitable for further mathematical treatment.

3. It cannot be computed for distribution with open-end classes.

only MOD which can be compute for open end classes is Quartile Deviation.

$$\text{Coeff of MD} = \frac{\text{MD about mean}}{\text{mean}} \times 100$$

or

$$\frac{\text{MD about Median}}{\text{Median}} \times 100$$

Quartile Deviation → Semi-inter quartile Range

Q_1 Q_2, Q_3

- Inter-quartile range = $Q_3 - Q_1$

- Quartile deviation, also called semi-inter-quartile range,

- Quartile Deviation (Q.D.) = $\dfrac{Q_3 - Q_1}{2}$
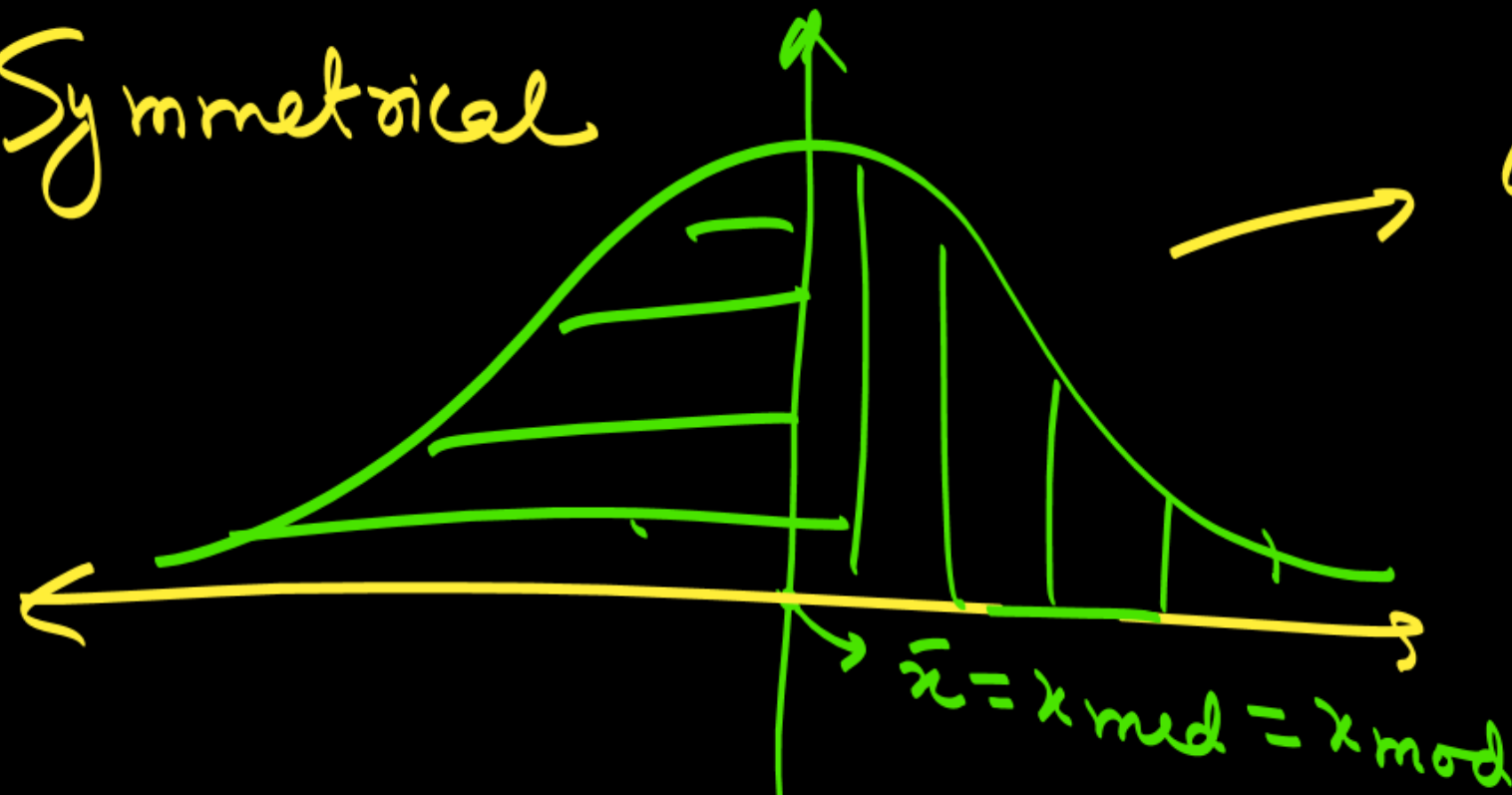
- The coefficient of quartile deviation is a relative measure of dispersion defined by

  Coefficient of Quartile Deviation = $\dfrac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

- Coefficient of quartile deviation is a pure number and can be used to compare two distributions expressed in different

Symmetrical

$$Coef. \text{ of } QD$$

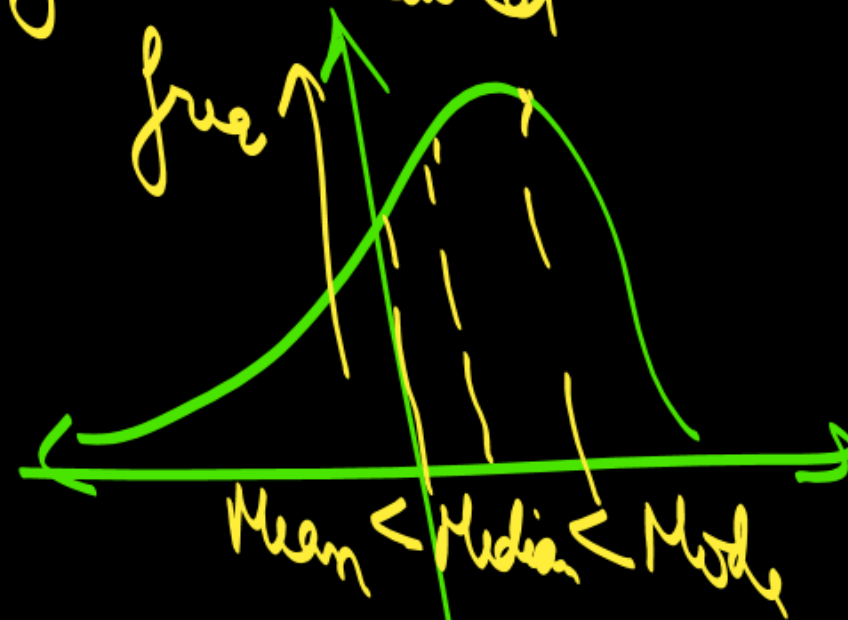$$= \frac{QD}{Median} \times 100$$

$\bar{x} = x\,med = x\,mod$

Skewed

Positively → ← Negatively Skewed
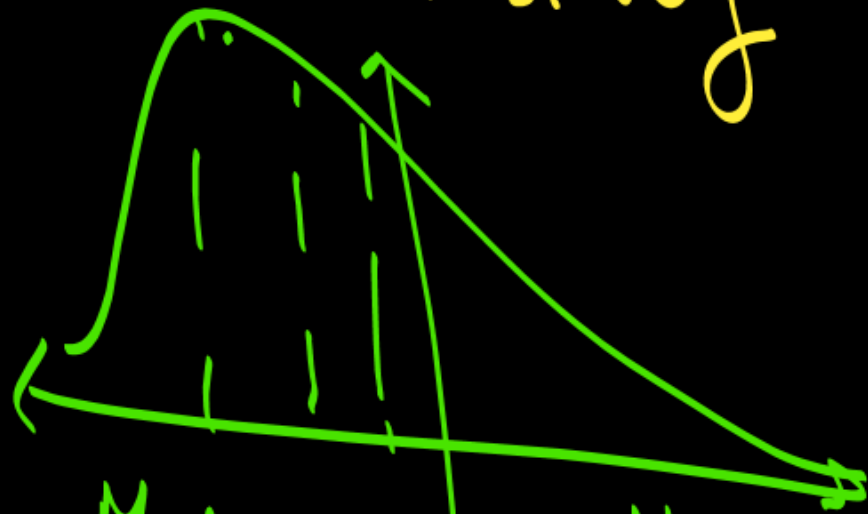
Mode < Median < Mean

Mean < Median < Mode

Properties

- Like other measures of dispersion, quartile deviation remains unaffected due to a change of origin but is affected in the same ratio due to change in scale.

$$y = a + bx$$

$$QD_y = |b| \times QD_x$$

Merits

- Quartile deviation provides the best measure of dispersion for open-end classification.

- It is also less affected due to sampling fluctuations

- Quartile deviation is useful specially when it is desired to study variability in the central half part of the data.

Demerits

- Quartile deviation is not based on all the observations.

- Quartile deviation is not suitable for further mathematical treatment.

- It is affected considerably by sampling fluctuations.
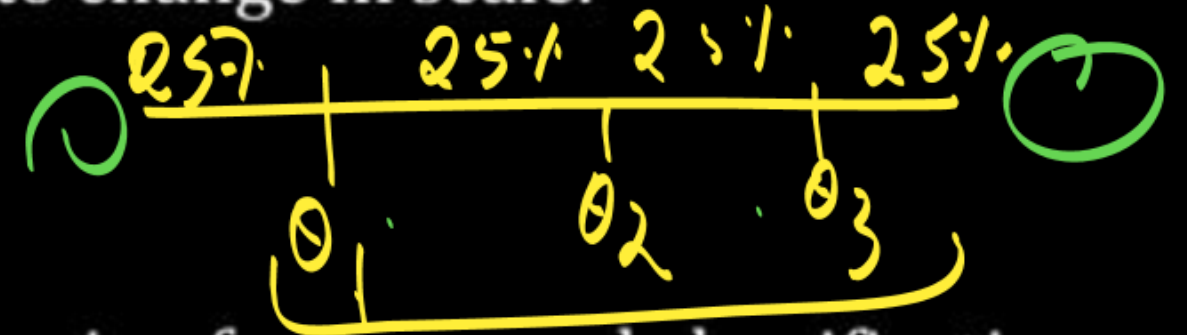
If x and y are related as $3x + 4y = 20$ and the quartile deviation of x is 12, then the quartile deviation of y is

A) 16

B) 14

C) 10

D) 9

$3x \oplus 4y = 20$

$QD_x = 12$

$QD_y = ?$

$3 \, QD_x = 4 \, QD_y$

$3(12) = 4 \, QD_y$

$QD_y = \dfrac{3 \times \cancel{12}^{3}}{\cancel{4}} = 9$

Measures of Dispersion

Chart

	Absolute	Relative	If y = a + bx
RANGE (R)	Range = Largest (L) – Smallest (S)	Co efficient of Range $= \frac{L-S}{L+S} \times 100$	$R_y = \lvert b \rvert \times R_x$
MEAN DEVIATION (M.D) about A	$M.\,D_A = \frac{1}{n}\sum \lvert x - A \rvert$	Co efficient of M.D from A $= \frac{M.D\ about\ A}{A} \times 100$	M.D $D_y = \lvert b \rvert \times$ M.D D_x
MEAN DEVIATION (M.D) about A.M ($\bar{x}$)	M.D about Mean $= \frac{1}{n}\sum \lvert x_i - \bar{X} \rvert$	Co efficient of M.D from A.M $= \frac{M.D\ about\ \bar{x}}{\bar{x}} \times 100$	M. $D_y = \lvert b \rvert \times$ M · D_x
MEAN DEVIATION (M.D) about Median	M.D about Median $= \frac{1}{n}\sum \lvert x_i - Median \rvert$	Co efficient of M.D from Median $= \frac{M.D\ about\ A}{A} \times 100$	M.D $D_y = \lvert b \rvert \times$ M · D_x

Standard Deviation

→ Best MOD

It is the positive square root of the arithmetic mean of the squares of deviations of the observations from their arithmetic mean.

Root mean squared deviation from mean

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

Coefficient of Variation

- The coefficient of variation,, is a relative measure of dispersion and is given by

Coefficient of Variation = $\dfrac{\sigma}{\bar{x}} \times 100$

- This is a pure number independent of the units of measurement and hence can be used to compare the variability of two distributions expressed in different units.

- A distribution for which the coefficient of variation is smaller is said to be less variable or more consistent, more uniform, more stable or more homogeneous.

- On the other hand, the distribution for which the coefficient of variation is greater is said to be more variable or less consistent, less uniform, less stable or less homogeneous.

The algebraic sum of 10 items about 8 is -5. Find its arithmetic mean. Also find coefficient of variation if standard deviation is 1.5.

Numerical
Question

→ Next
Class of chankyaniti

- Standard deviation is suitable for further mathematical treatment.

- Standard deviation is independent of change of origin but not of scale, if $y = a + bx$ for any two constants a and b, then
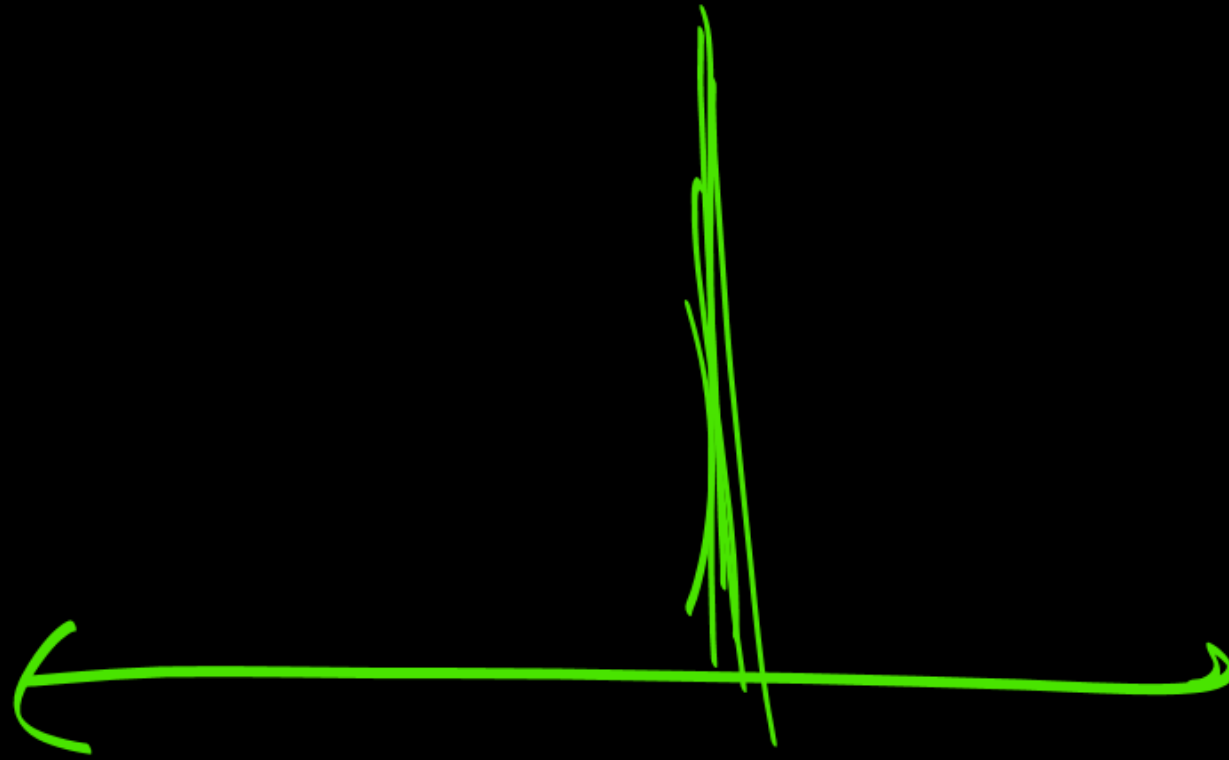
  SD of y =

$$y = a + b.x$$
$$SD_y = |b| \times SD_x$$

- If all the observations assumed by a variable are constant i.e., equal, then the SD is zero. This result applies all measure of dispersion

$$\rightarrow 1, 2, 3, 4, 5$$

- The standard deviation of first n natural numbers is $= \sqrt{\dfrac{n^2-1}{12}}$

$$\sigma = \sqrt{\frac{n^2-1}{12}} = \sqrt{\frac{5^2}{12}} = \sqrt{\frac{25-1}{12}} = \sqrt{2}$$

If all value are same
$\downarrow$

2, 2, 2,
$\downarrow$
Range = 0

MD = 0

QD = 0

SD = 0

Find SD of 1, 2, 3, 4, 5, 6, 7, 8, 9

A $\sqrt{\dfrac{20}{3}}$

B $\sqrt{\dfrac{81}{3}}$

C $\sqrt{\dfrac{20}{5}}$

D None of these

$$SD = \sqrt{\dfrac{9^2-1}{12}} = \sqrt{\dfrac{81-1}{12}} = \sqrt{\dfrac{80}{12}} = \sqrt{\dfrac{20}{3}}$$

- **If frequencies of all observations are same, count them once only.**

$$SD = \frac{Range}{2}$$

S.D. between 2 No.'s = $\frac{|a-b|}{2}$

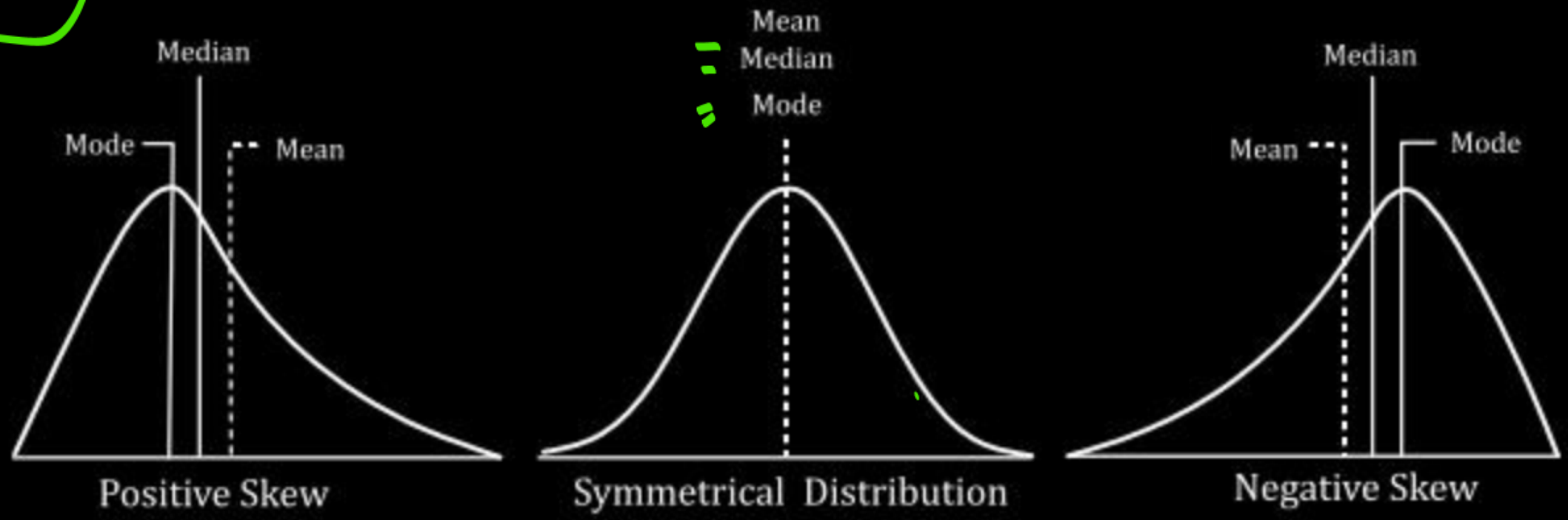$$SD = \frac{8 - 2}{2} = \frac{6}{2} = 3$$

$$SD = \frac{|2 - 8|}{2}$$

$$= \frac{|-6|}{2} = 3$$

- **Relationship between MD, QD & S.D.** (Symm.

$$\boxed{4S.D. = 5MD = 6QD}$$

Skewness

Mode — Median — Mean (Positive Skew)

Mean = Median = Mode (Symmetrical Distribution)

Mean — Median — Mode (Negative Skew)

For any two numbers SD is always

(A) Twice the range

(B) Half of the range

(C) Square of the range

(D) None of these

$$a, b$$

$$SD = \frac{|a-b|}{2} \rightarrow \frac{Range}{2}$$

The standard deviation of 10, 16, 10, 16, 10, 10, 16, 16 is

(A) 4

(B) 6

(C) 3

(D) 0

$$\frac{x_i \mid y}{10 \mid 4}$$
$$16 \mid 4$$

$$x_i \Rightarrow 10, 16$$

$$SD = \frac{16-10}{2} = \frac{6}{2} = 3$$

Standard Deviation

Variance = $(SD)^2$

	Absolute	Relative	If y = a + bx
Standard Deviation (σ)	$\sigma = \sqrt{\dfrac{\sum(x_i - \bar{X})^2}{n}}$ $\sigma = \sqrt{\dfrac{\sum x_i^2}{n} - \bar{X}^2}$	Co efficient of Variation = $\dfrac{\sigma}{x} \times 100$	$\sigma_y = \|b\| \times \sigma_x$
	Standard Deviation *for Two number* = 2 Standard Deviations for first n Natural numbers, $\sigma = \sqrt{\dfrac{n^2 - 1}{12}}$	Combined Standard Deviation, $\sigma_{12} = \sqrt{\dfrac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$ Where $d_1 = \bar{x}_1 - \bar{x}_{12}$, $d_2 = \bar{x}_2 - \bar{x}_{12}$	
Variance (σ^2)	Variance means Square of standard Deviation		

If all the observations are Increased by 10, then

add

$$y = x + 10$$

A SD would be increased by 10

B Mean deviation would be increased by 10

C Quartile deviation would increase by 10

D All these three remain unchanged.

$$MD_y = MD_x$$

$$QD_y = QD_x$$

$$SD_y = SD_x$$

If all the observations are, multiplied by 2, then

$$y = 2 \cdot x$$

A New SD would be also multiplied by 2

$$SD_y = 2 \cdot SD_x$$

B New SD would be half of the previous SD

C New SD would be increased by 2

D New SD would be increased by 2

The best statistical measure used for comparing two series is

A Mean absolute deviation

B Range

C Coefficient of variation

D Standard deviation
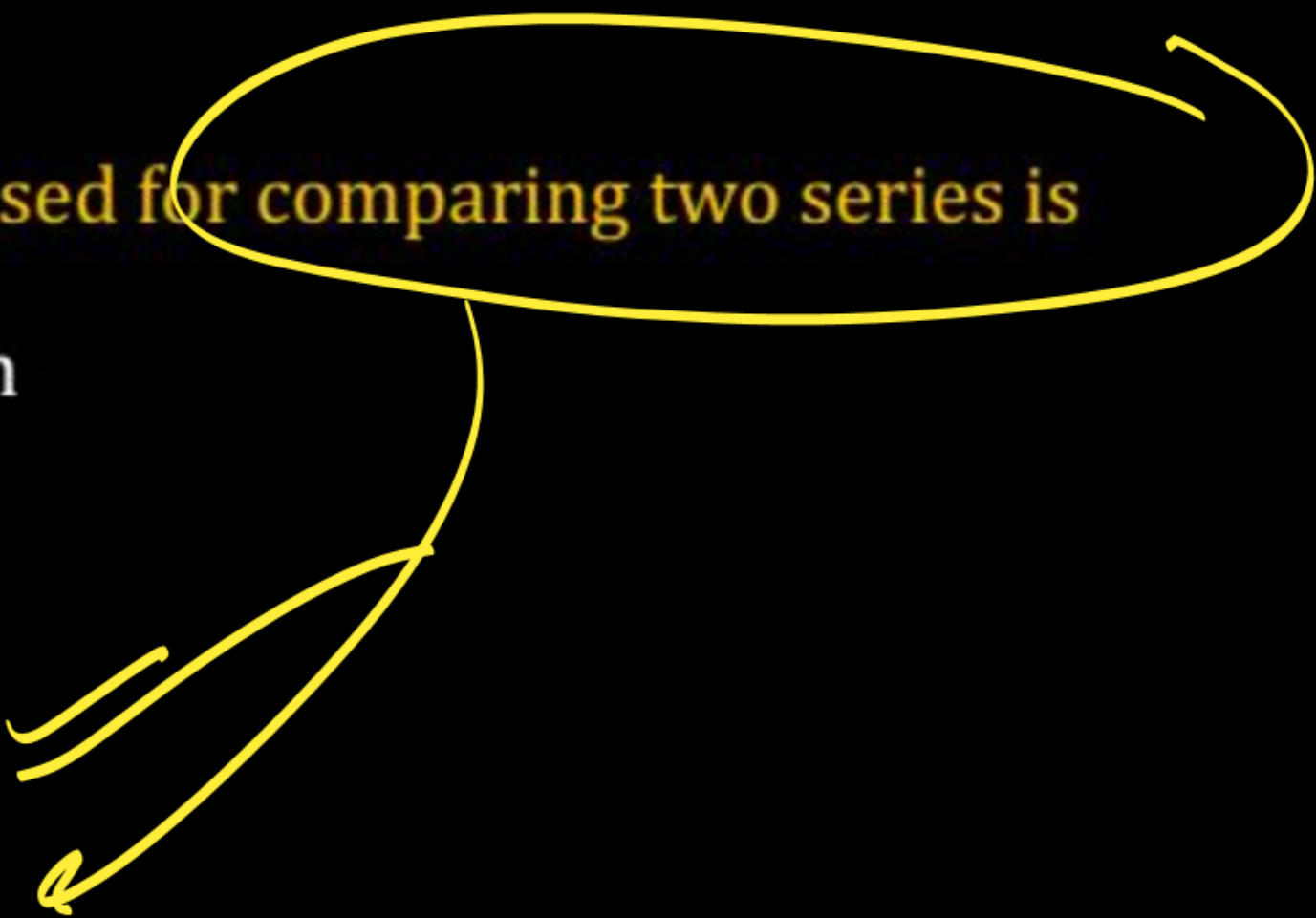
Which of the following is a relative measure of dispersion?

(A) Range

(B) Mean deviation

(C) Standard deviation

(D) Coefficient of quartile deviation

The standard deviation for the set of numbers 1, 4, 5, 7, 8, is 2.45 nearly. If 10 is added to each number then new standard deviation is

A) 24.45

B) 12.45

C) 2.45

D) 0.245

$x \rightarrow SD_x = 2.45$

$y = x + 10$

$SD_y = SD_x = 2.45$