"The degree to which numerical data tends to spread about an avg value is variation/dispersion"

Measures of Dispersion :-

good MOD should be/have :
- → simple to understand
- → easy to compute
- → rigidly defined
- → based on each & every item.
- → amendable to further algebraic treatment
- → sampling stability
- → not be affected by extreme items.

Methods of MOD :-

1. Range
2. Interquartile range & quartile deviation } Positional measures
3. Mean deviation / Avg dev
4. Standard deviation } Calculation measures
5. Lorenz curve → graphic method.

Absolute measure of variation
↳ same statistical unit
eg → ₹, Kg, tonnes etc

Relative measures of variation
↳ two sets of data expressed in diff units
eg → quintals of sugar & tonnes of sugarcane

→ relative dispersion is ratio of absolute dispersion to appropriate avg.

→ Called "Coefficient of dispersion" cuz 'coefficient' means a pure number which is unit less.

① Range

→ simplest method
→ diff b/w largest & smallest

$\boxed{\text{Range} = L - S}$ → absolute

→ $\boxed{\text{Coeff of Range} = \frac{L-S}{L+S}}$ → relative

→ In case of grouped data, either {lower limit of lowest class & higher limit of highest class} are taken OR {midpoint of lowest class & midpoint of highest class}

Merits :- simplest & easiest
- minimum time to calculate
- quick rather than accurate

Limitations :- Not based on each & every item
- shows fluctuations from sample to sample
- unreliable as guide to dispersion of values within the extremes

eg →

(A)	46	6	46	46	46
(B)	6	10	7	46	40
(C)	6	6	25	30	46

All have Range = 46 − 6 = 40

- Cannot be computed for open-end distribution

② ~~Inter~~ Interquartile Range / Quartile deviation

- includes mid 50 % of distribution
- ~~leaves~~ excludes starting & ending 25% (quarters)
- Diff b/w third quartile & first quartile

i.e $\boxed{\text{Interquartile range} = Q_3 - Q_1}$

- often reduced to semi-inter-quartile range, a.k.a. quartile deviation by dividing by 2.

i.e $\boxed{\text{Quartile deviation} = QD = \frac{Q_3 - Q_1}{2}}$ → absolute

- Q1 & Q3 are equidistant from Median (Q_2) i.e $Med - Q_1 = Q_3 - Med$
∴ Median ± Q.D. Covers 50% of observations.
& for asymmetrical distributions it is (approx 50 %)

- $\boxed{\text{Coeff of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}}$ → relative

$Q_1 = \text{Size of} \left(\frac{N+1}{4}\right)^{th} \text{term}$ [for odd]

$Q_1 = \text{Size of} \left(\frac{N}{4}\right)^{th} \text{term}$ [for even]

$Q_3 = \text{Size of } \frac{3}{4}(N+1)^{th} \text{ term}$ [for odd]

$Q_3 = \text{Size of } \frac{3}{4} N^{th} \text{ term}$ [for even]

For Continuous Series

$$\boxed{\begin{aligned} Q_1 &= L + \frac{\frac{N}{4} - P.C.f.}{f} \times i \\ Q_3 &= L + \frac{\frac{3N}{4} - P.C.f}{f} \times i \end{aligned}}$$

where, L = Lower limit of Class
P.C.f = Prev. Cummulative freq
f = freq
i = size of Class = Upper - Lower limit

Merits :-
- superior to range
- can be used for open-end distribution
- Not affected by extreme values
∴ useful for badly skewed distribution.

Limitations :-
- ignores 50% (first & last 25%) items
∴ doesn't depend upon every item
- not capable of mathematical manipulation (not ammendable)
- affected by sampling fluctuations.
- It is more as measure of partition than a measure of dispersion as it doesn't show scatter around an avg.

$\boxed{\text{Percentile Range} = P_{90} - P_{10}}$

$\boxed{\text{Semi percentile range} = \frac{P_{90} - P_{10}}{2}}$

③ Mean deviation / Avg deviation

- avg dist b/w items & the mean or median of series.
- deviations from median has more advantage cuz the sum of deviations from median is minimum when signs are ignored
But, mean is more frequently used in practice ∴ it is called 'mean' deviation.

For Individual Observations

$$\boxed{M.D. = \frac{\Sigma |D|}{N}}$$

$|D|$ is deviation from mean or median, ignoring sign.

$$\boxed{\text{Coeff of MD} = \frac{M.D}{\text{Median}} \text{ or } \frac{MD}{\text{Mean}}}$$

- For normal distribution, M.D ± mean or median will include 57.5 % of items.

$$\boxed{\text{Median} = Q_2 = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item [For odd] or size of } \left(\frac{N}{2}\right)^{th} \text{ item [For even]}}$$

For Discrete series

$$\boxed{M.D. = \frac{\Sigma f|D|}{N}}$$

For Continuous series
[Same formula] with $|D| = \frac{|m - \bar{X}_{\text{or Med}}|}{C}$
m = mid point of class, C = common factor

Mean, $\bar{X} = A + \frac{\Sigma fd}{N} \times C$

A = assumed mean.
f = freq, d = m − A
m = mid point of class
C = common factor
[d is with signs + AND −]

Short cut method

$$\boxed{MD = \frac{\Sigma mf_A - \Sigma mf_B - (\Sigma f_A - \Sigma f_B)\bar{X}_{\text{or med}}}{N}}$$

where f_A = freqs above avg
f_B = freqs below avg
m = mid points

Merits :-
- Simplicity, presentable to general public.
- based on each & every item, changes value if any item is changed
- Less affected by extreme items than standard deviation.

Limitations :-
- Signs are ignored, makes it non-algebraic
- less accuracy
- not capable of further algebraic treatment
- rarely used in sociological studies

Thus, overshadowed by the superior standard deviation.

④ Standard Deviation

- Intro by Karl Pearson, most imp & widely used.
- Also called Root-Mean Square deviation ∵ tis sqrt of means of squared deviations from mean.

Deviations from actual mean

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

Deviations from assumed mean

$$\boxed{\sigma = \sqrt{\frac{\Sigma(x-A)^2}{N} - \left(\frac{\Sigma(x-A)}{N}\right)^2}}$$

[or $x - A = d$]

For discrete series, $\boxed{\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}}}$ ↳ Actual mean

~~For continuous ser~~

$$\boxed{\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}}$$

where, $d = x - A$ ↳ Assumed mean

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times C$$

$d = \frac{x-A}{C}$, C = common factor

↳ step deviation method

↓

Same for Continous series except $d = \frac{m-A}{c}$, m = mid point

Mathematical Properties:-

① Combined S.D.

$$\boxed{\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1+N_2}}}$$

where $d_1 = \bar{X}_1 - \bar{X}_{12}$, $d_2 = \bar{X}_2 - \bar{X}_{12}$

$$\boxed{\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2}{N_1+N_2+N_3}}}$$

where $d_1 = \bar{X}_1 - \bar{X}_{123}$, $d_2 = \bar{X}_2 - \bar{X}_{123}$, $d_3 = \bar{X}_3 - \bar{X}_{123}$

Note → Similar for combined mean

$$\boxed{\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1+N_2}}$$

② S.D. of n natural number

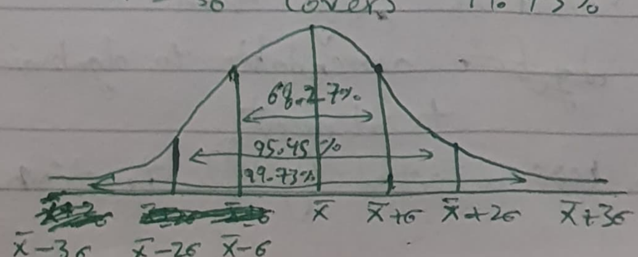
$$\boxed{\sigma = \sqrt{\frac{1}{12}(N^2-1)}}$$

③ Sum of deviations from mean is minimum i.e the sum deviations from any other value would've been greater than from mean ∴ mean is used

④ Mean $\pm 1\sigma$ covers 68.27%

$\bar{x} \pm 2\sigma$ covers 95.45%

$\bar{x} \pm 3\sigma$ covers 99.73%



~~Rela~~ Relation b/w MoD

$$\boxed{Q.D = \frac{2}{3}\sigma = 0.6745\sigma}$$

$$\boxed{M.D. = \frac{4}{5}\sigma = 0.7979\sigma}$$

$\bar{x} \pm Q D$ includes 50 %
$\bar{x} \pm M.D$ includes 57.51%
$\bar{x} \pm \sigma$ includes 68.27% (2/3 items)

Coefficient of Variation

i.e $\boxed{C.V. = \frac{\sigma}{\bar{X}} \times 100}$

Note → Could've used any MoD & any avg to find coeff of variability but almost always, SD is used along with arithematic mean.

Variance

$$\text{Variance} = \sigma^2 = \frac{\Sigma(x-\bar{X})^2}{N}$$

or $$= \left\{ \frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2 \right\} \times C^2$$

$d = \frac{x - A}{C}$; C = common factor

Merits :-

- Based on every item
- ~~algebric~~ amendable to algebraic treatment
- less affected by fluctuations of ~~samplen~~ sampling.
- possible to calculate combined SD
- comparison of two or more distribution is possible due to coeff of variation.
- used further in computing skewness, correlation etc, helpful in sampling, provides unit of measurement for normal distribution.

Limitations

- Difficult to compute but highly accurate
- Gives more weight to extreme items & less to items near mean ~~because~~ cuz of squaring
  eg → deviations 2 & 8 are in ration 1:4 but their squares 4 & 64 are in ratio 1:16.

Diff b/w MD & SD

- Algebraic signs are ignored in MD while they're taken in account for SD
- MD can be ~~comput~~ computed from either mean or median while SD is always computed from mean.

8 7 WHICH MEASURE OF DISPERSION TO USE

Like measures of central value, in case of measures of variation also, the choice of a suitable measure depends on the following two factors:

1 *The type of data available.* If they are few in numbers, or *contain extreme values, avoid the standard deviation* *If they are* generally skewed, avoid the mean deviation as well. If they have gaps around the quartiles, the quartile deviation should be avoided. If there are open-end classes, the quartile measure of dispersion should be preferred.

2 *The purpose of investigation.* In an elementary treatment of statistical series in which a measure of variability is desired only for itself, any of the three measures, namely, range, quartile deviation and average deviation, would be acceptable Probably the average deviation would be better However, in usual practice, the measure of variability is employed in future statistical analysis. For such a purpose, the standard deviation by far is the most popularly used It is free from those defects from which other measures suffer. It lends itself to the analysis of variability in terms of normal curve of error.* *Practically all advanced statistical* methods deal with variability and centre around the standard deviation. Hence unless the circumstances warrant the use of any other measure, we should make use of standard deviation for measuring variability.

Measures of Dispersion (Module 4)

The measure of central tendency do exhibit one of the important characteristic of a distribution, yet they fail to give any idea as to how the individual values ~~deform~~ differ from the central values i.e whether ~~ty~~ they are closely packed around the central value or widely scattered away from it. Two distribution may have the same mean & same total frequency, for example :-

A	B	C
100	102	1
100	103	2
100	98	3
100	97	490
100	100	4
500	500	500

~~mean~~, $\bar{x}_A = 100$, $\bar{x}_B = 100$ $\bar{x}_C = 100$

Definition - Dispersion is the measure of variation of items
by Bowley

Definition - The degree of which numerical data tends to spread
by Spiegal about an average value is called dispersion of the data.
(variation)

Methods to find measures of dispersion

① Range - Range is given by the difference between the greatest and least value in the distribution

share price fluctuation
weather forecast

i.e $\boxed{\text{Range} = \text{Max value} - \text{Min value}}$

② Quartile deviation - $\boxed{\dfrac{Q_3 - Q_1}{2} = \text{Quartile deviation}}$

③ mean-deviation

$$\text{mean deviation} = \frac{\sum f_i\,|(x_i - A_i)|}{\sum f_i = N}$$

④ Standard deviation - (i) $SD = \sigma = \sqrt{\dfrac{\sum f_i (x_i - A)^2}{\sum f_i}}$

(ii) Short Cut Method

$$\sigma = \sqrt{\frac{\sum f_i d_i^2}{\sum f_i} - \left(\frac{\sum f_i d_i}{\sum f_i}\right)^2}$$

(iii) Step deviation

$$\sigma = h \times \sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i}\right)^2}$$

where, $u_i = \dfrac{x - a}{h}$

$$\text{variance} = \sigma^2 = \frac{\sum f_i (x_i - A)^2}{\sum f_i}$$

⑤ Coefficient of variation

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$ → for %

List of Formulae

Individual Observations	*Discrete & Continuous Series*
Range $\text{Range}=L-S$ $\text{Coeff. of Range}=\frac{L-S}{L+S}$	(Same as on the left) But L, i.e., largest value, will be the *upper limit of the highest class and S* will be the lower limit of the lowest class
Quartile Deviation $Q.D.=\frac{Q_3-Q_1}{2}$ $\text{Coeff of } Q.D.=\frac{Q_3-Q_1}{Q_3+Q_1}$	(Same as on the left)
Mean Deviation $M.D.=\frac{\Sigma\mid D\mid}{N}$ $\text{Coeff. of } M.D.=\frac{M.D.}{\text{Median}}$ or $\frac{M.D.}{\text{Mean}}$ (if deviations are taken from mean)	$M.D.=\frac{\Sigma f\mid D\mid}{N}$ $\text{Coeff of } M.D.=\frac{M.D.}{\text{Median}}$ or $\frac{M.D.}{\text{Mean}}$ (if deviations are taken from mean)
Standard Deviation *Actual Mean Method* $\sigma=\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$	*Actual Mean Method* $\sigma=\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$
Assumed Mean Method $\sigma=\sqrt{\frac{\Sigma d^2}{N}-\left(\frac{\Sigma d}{N}\right)^2}$ *Step Deviation method* $\sigma=\sqrt{\frac{\Sigma d'^2}{N}-\left(\frac{\Sigma d'}{N}\right)^2}\times C$ $C.V.=\frac{\sigma}{\bar{X}}\times 100$	*Assumed Mean Method* $\sigma=\sqrt{\frac{\Sigma fd^2}{N}-\left(\frac{\Sigma fd}{N}\right)^2}$ *Step Deviation Method* $\sigma=\sqrt{\frac{\Sigma fd'^2}{N}-\left(\frac{\Sigma fd'}{N}\right)^2}\times C$ $C.V.=\frac{\sigma}{\bar{X}}\times 100$

Combined Standard Deviation

$$\sigma_{12}=\sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_1d_1^2+N_2d_2^2}{N_1+N_2}}$$

Where $d_1=(\bar{X}_1-\bar{X}_{12})$ and $d_2=(\bar{X}_2-\bar{X}_{12})$