Measures of Central Tendency:-

- Measures of central tendency yield info. about "particular places (or) locations in a group of no's".
- A single no. to describe the characteristics of a set of a data.

Summary statistics:-

- Central tendency (or) measures of location
  - Arithmetic Mean
  - Weighted Mean
  - Median
  - Percentile
- Dispersion
  - skewness
  - kurtosis
  - Range
  - Interquartile range
  - Variance
  - Standard score
  - Coefficient of Variation.

Arithmetic Mean:-

- Commonly called 'the mean', it is the average of a group of no's.
- Applicable for interval & ratio data
- Not applicable for nominal & ordinal data.
- Affected by each value in the dataset, including extreme values.
- Computed as $\frac{\sum_{i=1}^{n} x_i}{n}$ where n = #values (or) rows in the dataset.

Four levels of Measurement:-

Nominal data:- data that is classified into categ
& cannot be arranged in any particular order.
ex:- color, gender etc.

Ordinal data:- Involves data arranged in some orde
but the diff's blw data values cannot be determin
(or) are meaningless.
Ex:- Quality of service on a scale of 1 to 5
with 1 indicating very dissatisfied & 5 indicating very
satisfied.

Interval data:- similar to ordinal data, with additional
property that meaningful amt's of diff's blw data
values can be determined. There is no natural zero
point.
ex:- Temperature on fahrenheit scale

Ratio data:- Interval data with an inherent zero
starting point. Differences & ratios are meaningful
for this level of measurement.
Ex:- Monthly income, distance travelled in a week
etc.

Population mean: Denoted by μ.

$$\mu = \frac{\sum X}{N}$$, where $X \rightarrow$ item in the population

$N \rightarrow$ # items in the population

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population = [24, 13, 19, 26, 11]

Ex: $\mu = \frac{24+13+19+26+11}{5} = \frac{93}{5} = 18.6$

Sample Mean: Calculated for a sample, same as population mean. Denoted as $\bar{x}$.

$$\bar{x} = \frac{\sum x}{n}$$

$n \rightarrow$ # items in the sample

$x \rightarrow$ items in the sample.

Mean of Grouped data:

- Weighted average of class midpoints.
- class frequencies are the weights.

$$\mu = \frac{\sum_i f_i x_i}{\sum_i f_i} \Rightarrow \mu = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n}$$

where $f_i \rightarrow$ frequency of i^{th} class interval

$x_i \rightarrow$ midpoint of i^{th} class interval

Calculating mean of grouped data example:

in next page.

Class Interval	Frequency (f)	class Mid.pt. (X)	fX
20 – 30	6	25	150
30 under 40	18	35	630
40 under 50	11	45	495
50 under 60	11	55	605
60 under 70	3	65	195
70 under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fX}{\sum f} = \frac{2150}{50} = 43.0$$

Weighted average:-

- Sometimes we wish to average no's, but we want to assign more importance, or weight, to some of the no's.
- The average you need is the weighted average.

$$\text{weighted average} = \frac{\sum xw}{\sum w}$$

where x is a data value & w is the weight assigned to that data value. The sum is taken over all data values

Ex:- Suppose your midterm test score is 83 & your final exam score is 95. Using weights of 40% for midterm & 60% for final exam, compute the weighted average of your scores. If the min. avg for an A is 90 will you earn A?

Ans: weighted average = $\frac{83(0.4)+95(0.6)}{0.4+0.6}$

$= \frac{33.2+57}{1} = 90.2$ → You'll earn A!

Median:-

- Middle value in an Ordered array of no's.
- Applicable for Ordinal, interval & ratio data.
- Not applicable for nominal data.
- Unaffected by extremely large & extremely small values

Median: Computational procedure:-

- First Procedure
  - Arrange the observations in an Ordered array.
  - If there are odd no. of terms, the median is the middle term of the Ordered array.
  - If there is an even no. of terms, the median is the avg. of the middle two terms.
- Second Procedure
  - The median's position in an Ordered array is given by $(n+1)/2$.

Median: Example with odd # terms:-

Ordered array:-

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

terms = 17 (odd)

Position of median = $(17+1)/2 = 9$ ∴ Median = 15

- If 22 is replaced by 100, median is 15
- If 3 is replaced by -103, median is 15.

Median: Example with even # terms

Ordered array:-

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

- # terms = 16 ⇒ Position of median = $(16+1)/2 = 8.5$
- The median is blw 8th & 9th terms, 14.5
- If 21 is replaced by 100, median is 14.5
- If 3 is replaced by -88, median is 14.5

Median of Grouped data:-

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

where

L: lower limit of median class

N: total # frequencies

cf_p: cumulative frequency of class preceding median class

f_{med}: frequency of median class

W: width of median class.

Ex:-

Class Interval	Frequency	Cumulative Frequency	
20 under 30	6	6	
30 under 40	18	24	
40 under 50	11	35	→ Median Class
50 under 60	11	46	
60 under 70	3	49	
70 under 80	1	50	
	N=50		

median class: 40 under 50 since $\frac{50}{2} = 25$ is there in that class.

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f_{med}}(w) = 40 + \frac{\frac{50}{2} - 24}{11}(10)$$

$$= 40.909$$

Mode:-

- The most frequently occuring value in a dataset.
- Applicable to all levels of data measurement (nominal, ordinal, interval & data).
- Bimodal -- Datasets that have two modes
- MultiModal -- Datasets that contain > 2 modes

Ex:- Given data is

35, 41, 44, 45, 37, 41, 44, 46, 37, 43, 44, 46, 39, 43, 44, 46, 40, 43, 44, 46, 40, 43, 45, 48

Here Mode is 44.

Mode of Grouped data:-

- Midpt of Modal class
- Modal class has the greatest frequency

Class Interval	Frequency
20 under 30	6
30 under 40	18
40 under 50	11
50 under 60	11
60 under 70	3
70 under 80	1

→ Modal class

$$\text{Mode} = L_{MO} + \left(\frac{d_i}{d_1 + d_2}\right)w$$

$$= 30 + \left(\frac{(18-2)}{(18-2)+(18-11)}\right)10$$

$$= 36.31$$

Percentiles:-

- Measures of central tendency that divide a group of data into 100 parts.
- Ex:- 90th percentile indicates that atmost 90% of the data lie below it, & at least 10% of the data lie above it.
- The median & 50th percentile have the same value.
- Applicable for ordinal, interval & ratio data.
- Not Applicable for nominal data.

- Computational procedure:-
  - Organise the data into an ascending ordered array.
  - calculate the p^{th} percentile location:

$$i = \frac{P}{100}(n)$$

  - Determine the percentile's location & it's value:
    - If 'i' is a whole no. the percentile is the avg. of the values at the 'i' & 'i+1' positions.
    - If 'i' is not a whole no. the percentile is at the $(i+1)^{th}$ position in the Ordered array.

Ex:- Raw data: 14, 12, 19, 23, 5, 13, 28, 17

Ordered array: 5, 12, 13, 14, 17, 19, 23, 28

location of 30^{th} percentile:

$$i = \frac{30}{100}(8) = 2.4$$

∴ i is not whole no. the percentile is at 3^{rd} position in the Ordered array i.e; 13.

Dispersion:-

- Measures of variability describe the spread (or) the dispersion of a set of data.
- Tells about → Reliability of measure of central tendency.
- Used → To compare the dispersion of various samples.

~~Variability~~ Variability:-

~~No variability~~

~~3 3 3 3~~

No variability in Cash flow

3 3 3 3

Mean

3

Variability in Cashflow

2 4 4 2

3

- So mean is not a good measure of the variability.

Measures of Variability (or) dispersion:-

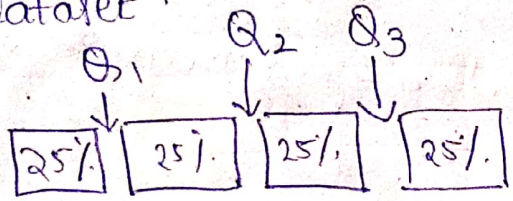
- Range
- Inter quartile range.
- Mean Absolute deviation
- Variance
- Standard deviation
- Z scores
- Coefficient of variation.

Range - Ungrouped data:-

- The diff. btw the largest & the smallest values in a set of data.
- Simple to compute & ignores all data pts. except the two extremes.

Quartiles:-

- Measures of central tendency that divide a group of data into four subgroups.
- Q_1: 25% of the dataset is below the first quartile
- Q_2: 50% of the dataset is below the second quartile
- Q_3: 75% of the dataset is below the third quartile.
- Q_1 is equal to the 25^{th} percentile
- Q_2 is located at 50^{th} percentile & equals the median.
- Q_3 is equal to the 75^{th} percentile.
- Quartile values are not necessarily members of the dataset.



Ex:- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

- Q_1 $\quad i = \frac{25}{100}(8) = 2$, $Q_1 = \frac{109 + 114}{2} = 111.5$
- Q_2 $\quad i = \frac{50}{100}(8) = 4$, $Q_2 = \frac{116 + 121}{2} = 118.5$
- Q_3 $\quad i = \frac{75}{100}(8) = 6$, $Q_3 = \frac{122 + 125}{2} = 123.5$

Interquartile range:-

- Range of values blw the first & third quartiles.
- Range of the "middle half"

- Less influenced by extremes.

Interquartile range = $Q_3 - Q_1$

Deviation from the Mean:-

- Dataset: 5, 9, 16, 17, 18
- Mean: $\mu = \frac{\Sigma X}{N} = \frac{65}{5} = 13$
- Deviations from the mean: $-8, -4, 3, 4, 5$

Mean Absolute Deviation:-

- Average of the absolute deviations from the mean.

X	$x-\mu$	$\lvert x-\mu \rvert$
5	-8	8
9	-4	4
16	3	3
17	4	4
18	5	5
	0	24

$$M.A.D = \frac{\Sigma |x-\mu|}{N}$$

$$= \frac{24}{5} = 4.8$$

Population variance:-

Avg. of the squared deviations from the arithmetic mean.

X	$x-\mu$	$(x-\mu)^2$
5	-8	64
9	-4	16
16	3	9
17	4	16
18	5	25
	0	130

$$\sigma^2 = \frac{\Sigma (x-\mu)^2}{N}$$

$$= \frac{130}{5} = 26.0$$

Population Standard Variance:-

- Square root of the Variance

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	3	9
17	4	16
18	5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5} = 26.0$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{26.0} = 5.1$$

Sample Variance:-

- Average of the Squared deviations from arithmetic mean

X	$X-\mu$	$(X-\mu)^2$
2398	625	390625
1844	71	5041
1539	-234	54756
1311	-462	213444
7092	0	663866

$$S^2 = \frac{\sum (X-\bar{X})^2}{n-1}$$

$$= \frac{663866}{3}$$

$$= 221288.67$$

n-1 because we lost one degree of freedom due to mean.

Uses of Standard deviation:-

- Indicator of financial risk.
- Quality Control

- Construction of quality control charts.
- process capability studies.

- Comparing populations
  - household incomes in two cities
  - employee absenteeism at two plants

Sample Standard Deviation:

- square root of the Sample Variance

X	X - μ	$(X-\mu)^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

$$= \frac{663,866}{3} = 221,288.67$$

$$S = \sqrt{S^2}$$

$$= \sqrt{221,288.67}$$

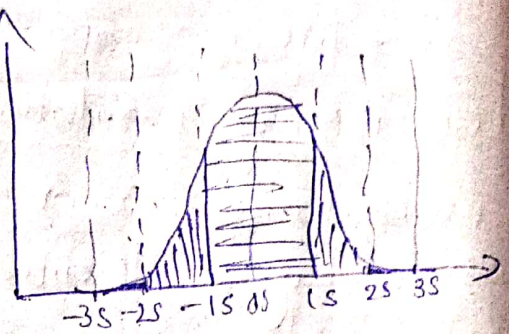
$$S = 470.41$$

Std. dev. as indicator of financial risk:-

	Annualized Rate of return	
Financial Security	μ	σ
A	15%	3%
B	15%	7%

The Empirical Rule :~:> If the histogram is bell shap

- Approximately 68%, 95%, 99.7% of all observations fall within 'one' 'two' 'three' standard deviations of the mean respectively.



- Data are normally distributed (or approximately normal)

Distance from the Mean	Percentage of values Falling within distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

Chebyshev's Theorem :- Not often used bcoz interval is very wide.

- A more general interpretation of the std. deviation is derived from Chebyshev's theorem, which applies to all shapes of histograms (not just bell shaped).
- The proportion of observations in any sample that lie within K std. dev's of the mean is at least,

$\boxed{1 - \frac{1}{K^2}}$ for $K > 1$.

For K = 2 (say) the theorem states that at least 3/4 of all observations lie within 2 std.dev's of the mean. This is a "lower bound" Compared to empirical rule's approximation (95%).

Coefficient of Variation:

- Ratio of the std. dev to the mean, expressed as a percentage.
- Measurement of relative dispersion

$$C.V = \frac{\sigma}{\mu}(100)$$

Ex:

$\mu_1 = 29$

$\sigma_1 = 4.6$

$$C.V_1 = \frac{\sigma_1}{\mu_1}(100)$$

$$= \frac{4.6}{29}(100)$$

$$= 15.86$$

$\mu_2 = 84$

$\sigma_2 = 10$

$$C.V_2 = \frac{\sigma_1}{\mu_2}(100)$$

$$= \frac{10}{84}(100)$$

$$= 11.90$$

- we use Coefficient of variation when mean Suggests one distribution & std. dev suggests another distribution.
- Lower the Coefficient of Variation, the better.

Variance & Std. dev of Grouped Data:-

Population

$$\sigma^2 = \frac{\sum f(M-\mu)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

Sample

$$s^2 = \frac{\sum f(M-\bar{x})^2}{n-1}$$

$$s = \sqrt{s^2}$$

Population Variance & Std. dev of Grouped data

(mu = 43)

Class Interval	f	M	fM	M-μ	$(M-\mu)^2$	$f(M-\mu)^2$
20 under 30	6	25	150	-18	324	1944
30 under 40	18	35	630	-8	64	1152
40 under 50	11	45	495	2	4	44
50 under 60	11	55	605	12	144	1584
60 under 70	3	65	195	22	484	1452
70 under 80	1	75	75	32	1024	1024
	50		2150			7200

$$\sigma^2 = \frac{\sum f(M-\mu)^2}{N} = \frac{7200}{50} = 144$$

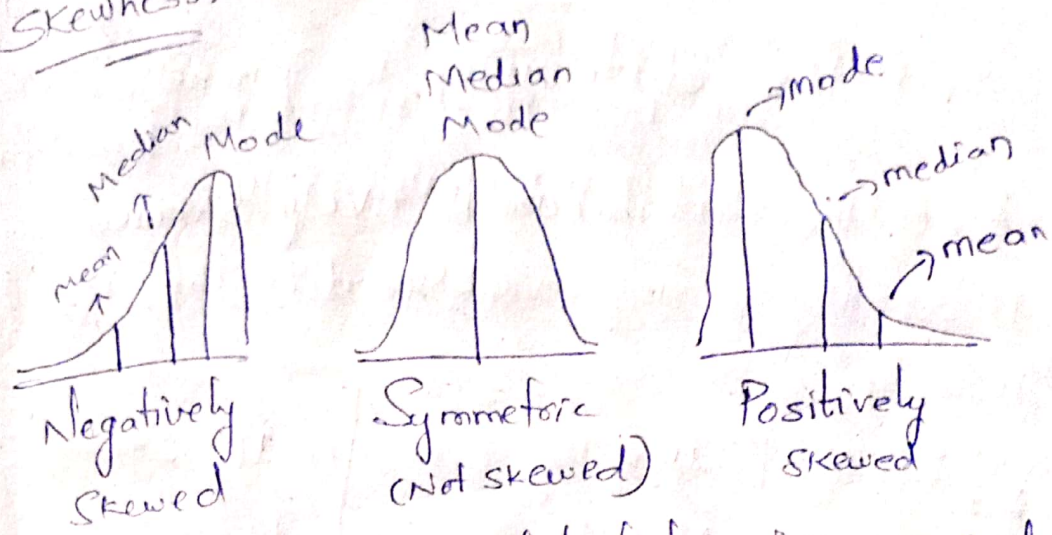
$$\Rightarrow \sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

Measures of Shape:-

- Skewness
  - Absence of symmetry
  - extreme values in one side of a distribution
- Kurtosis
  - Peakedness of a distribution
  - Leptokurtic: high & thin
  - Mesokurtic: normal shape
  - Platykurtic: flat & spread out
- Box & Whisker plots
  - Graphic display of a distribution

- Reveals skewness.

Skewness:-



Negatively Skewed | Symmetric (Not skewed) | Positively Skewed

- The skewness of a distribution is measured by comparing the relative positions of the mean, median & mode.
  - Distribution is Symmetrical
    - Mean = Median = Mode
  - Distribution skewed right
    - Median lies b/w mode & mean, & mode is less than mean.
  - Distribution skewed left
    - Median lies b/w mode & mean, & mode is greater than mean.

Coefficient of Skewness:-

- Summary measure for skewness

$$S = \frac{3(\mu - Md)}{\sigma}$$

- If $S<0$, the distribution is negatively skewed (skewed to the left)

- If $S=0$, the distribution is Symmetric (not skewed)

- If $S>0$, the distribution is +vely skewed (skewed to the right)

Ex:-

$\mu_1 = 23$

$M_{d_1} = 26$

$\sigma_1 = 12.3$

$S_1 = \frac{3(\mu - M_{d_1})}{\sigma_1}$

$= \frac{3(23-26)}{12.3}$

$= -0.73$

$\mu_2 = 26$

$M_{d_2} = 26$

$\sigma_2 = 12.3$

$S_2 = \frac{3(\mu - M_{d_2})}{\sigma_2}$

$= \frac{3(26-26)}{12.3}$

$= 0$

$\mu_3 = 29$

$M_{d_3} = 26$

$\sigma_3 = 12.3$

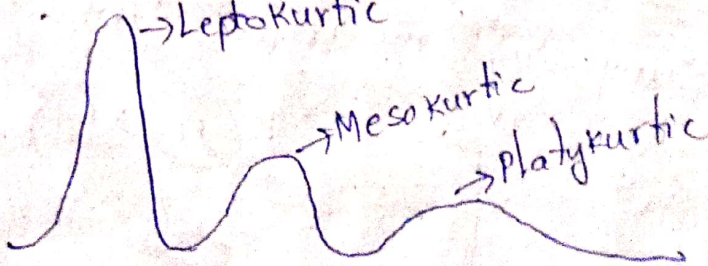
$S_3 = \frac{3(\mu - M_{d_3})}{\sigma_3}$

$= \frac{3(29-26)}{12.3}$

$= +0.73$

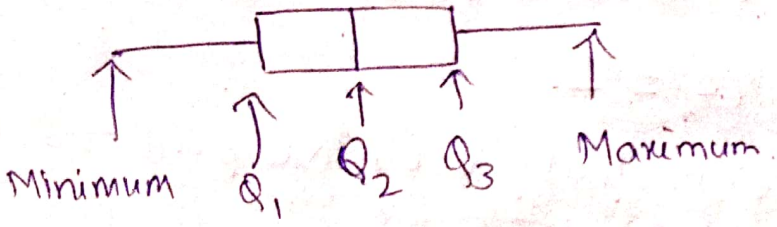
Kurtosis:-

- Peakedness of a distribution:-
  - Leptokurtic: high & thin
  - Mesokurtic: normal in shape
  - Platykurtic: flat & Spread out



Box & Whisker Plot:-

- Five specific values are used:
  - Median, Q_2
  - First Quartile, Q_1
  - Third Quartile, Q_3
  - Minimum value in the dataset
  - Maximum value in the dataset



Skewness: Box & Whisker Plots and Coefficient of Skewness:-