Statistics

Statistical Description of Data

'Statistics' is derived from –
- Latin word 'status'
- Italian word 'statista'
- German word 'statistik'
- French word 'statistique'.

Statistics defined in Plural sense –
Statistics is a collection of data for some specific purpose.

Singular sense –
Statistics is a collection, presentation & analyzing of data.
Statistics is a science of data collection, presentation and analysing.

Applications of Statistics

- Economics
- Business management
- Commerce & industry

* Economics – • many branches
• Econometrics
• Socio-economic survey
• Regression analysis.

* Business management – • How managers work, • making strategies
• statistical decision

Commerce and Industry → • Getting an edge in competitive business environment.

• To maximise profit
• Various tools used in commerce and industry of statistics

Limitations

• Study of quantitative data only.
• Study of aggregates only
• Homogeneity requirement
• Results may wrong
• used only by experts

Collection of Data

• On the basis of nature of data

• Quantitative data
(Numbers / Numeric)
ex→ Height, weight, no. of accidents etc.
⇓
measured as variables
→ Discrete → Constant • Fixed value • For any scale same value
→ Continuous → interval / approx.

• Qualitative data
(Characteristics / attributes).
ex→ Honesty, integrity, bravery etc.
⇓
measured as attributes.

Source of data

Primary source (collected by own or any agent originally).

Secondary source (collected data by other source).

4 ways to collect data

1) Interview
- Direct → Natural calamity
- Indirect → manmade accident
- Telephonic → podcast (fastest method)

2) mailed questionnaire ex-(Google form) — (max. non response)

(3) Observation — PT teacher

(4) Questionnaires filled and sent by enumerators — Aadhar card.

Sources of secondary data.

- International sources like WHO, ILO, WTO etc.
- Government sources like ministry of food etc
- Private and quasi-government sources like NCERT etc.
- unpublished source of various institutions.

★ Scrutiny of data → For quality check.

Presentation of data

- Textual / Discriptive
- Tabular method
- Graphical / Diagramatic method

★ Classification / Organising

It can be defined as the process of arranging data on the basis of characteristic into the number of groups or classes according to the similarity of observation.

Objectives –

- Simplification and briefness
- Comparability
- Statistical Analysis
- makes data more understable.

Types

1) Data chronological or Temporalor (Time series (Years)

2) Geographical or spatial Series or Data varying over the SPACE (States).

★ Frequency data → Quantity / Qualitative.

- Number of times a data value occurs.

★ Non-Frequency data → Geographical / Time Series.

- Single data value.

Graphical presentation of data

Tabular Presentation

Table no :- 01
Table name :- P/L balance Sheet

Row's head → column's head

Stub	Caption				Row's total
	Box Head (unit defined)				
	Column Head 1		Sub-Head 2		
Stub entries	Sub Col Head 1	Col Head 2	3	4	
1					
2					
-		Column total			

Rows

Body (main part)

Source :- ---

Footnote :- Abbreviation.

- Caption ⇒ The upper part of a table that describes the column or sub-column.
- Stub ⇒ Left part of the table describing the rows.
- Box head ⇒ Unit of measurement shown in box head.

Diagramatic Presentation

(i) Line diagram / Historiagram
Time – x axis
Fluctuation are wide – use Logarithm (Log chart / Ratio chart)

★ multiple line chart (multiple data in one chart) (same unit)
★ multiple axis chart (multiple data in one chart with diff units)

(ii) Bar diagram

→ Vertical Bar diagram : · Time Series · Quantitative
→ Horizontal bar diagram :- · Qualitative
· Data varying over space (geographical)

→ Multiple or Grouped Bar diagram :- 2011 (Feb, march, April)
→ Component or Sub-divided Bar diagram :- Jan (Potato, Tomato, Onion sale)

→ Dividend Bar diagram / % Percentage Bar diagram
Compare diff. components, comparing diff. components with table value.

(iii) Pie diagram (comparing different components to total value)

$$\text{Angle of component} = \frac{\text{Value of component}}{\text{Total value}} \times 100$$

Impt.- Best presentation method – Tabular
Hidden trend – Diagrammatic
most accurate - Tabular.

Frequency Distribution of data

Types

1) Grouped – making classes
2) ungrouped – data is in discrete variables (no classes)

- mutually exclusive ⇒ one value included in one class only.

★ Range = Largest value – Smallest value.

★ No. of class = $\frac{\text{Range}}{\text{Class length}}$.

★ No. of class interval × class lengths ≅ Range.

★ Class interval = Lower class limit – upper class limit
(LCL – UCL)

class length = 49 – 45 = 4

★★ For continuous variables, class limit and class boundaries are same. It is mutually exclusive.
Mutually exclusive classification that exclude the upper class limit.

Mutually inclusive classification that include both the class limits.

★ To convert class limit into class boundary :-

$$m = \frac{\text{UCL of a class} - \text{LCL of a class}}{2}.$$

$UCB = UCL + m$
$LCB = LCL + m$

★★ Discrete variable → ~~Ungrouped~~ frequency distribution → class interval → mutually inclusive → Simple frequency distribution

★ midpoint / mid value / class mark = $\frac{LCL + UCL}{2}$ or $\frac{LCB + UCB}{2}$

Cumulative frequency.
→ Less than
→ more than

- Frequency density of a class interval = $\frac{\text{Frequency of class interval}}{\text{class length}}$

- Relative Density = $\frac{\text{Frequency}}{\text{Total Frequency}}$.

- Percentage frequency = $\frac{\text{Frequency}}{\text{total frequency}} \times 100$

Sum of Relative frequency = 1 [lies between 0 and 1]

Sum of % frequency = 100

Graphical Representation of a Frequency distribution

(1) Histogram or Area diagram. (class boundaries)
⇓
use to find $\boxed{\text{MODE}}$

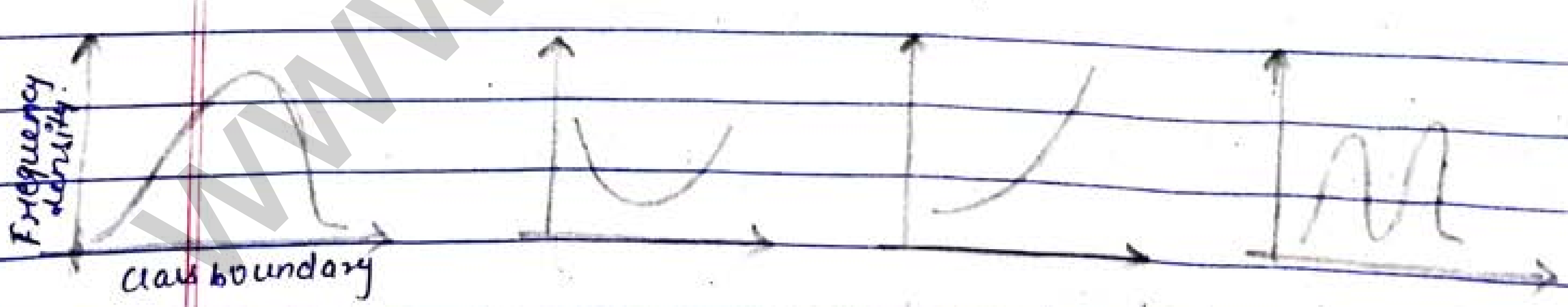
(2) Frequency polygon (joining the mid points of histogram and ends).

(3) Ogives / Cumulative Frequency curve
⇓
Gives the value of $\boxed{\text{MEDIAN}}$, $\boxed{\text{QUARTILES}}$.

(4) Frequency Curve. (It is a limiting form of a histogram or frequency polygon).

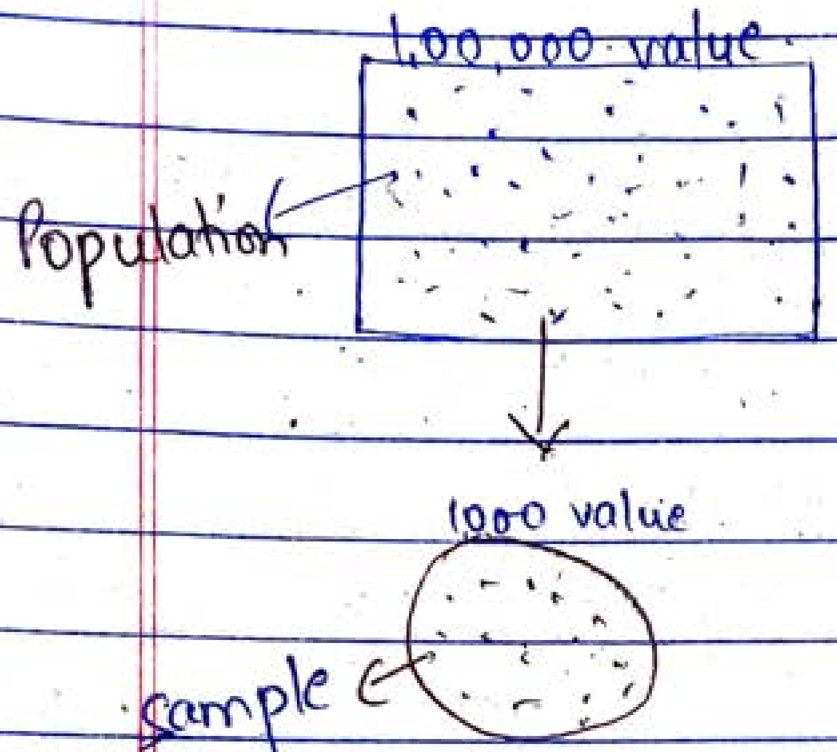
Four types of Frequency curve.

(a) Bell shaped curve
(b) U-shaped curve
(c) J-shaped curve
(d) mixed curve.



- Commonly used
ex: Traffic, Distribution of profit.

Sampling



Population → Population parameter
↓
sample → Sample statistic
↓
Hypothesis test.

* **Population** :- aggregate of all the units under consideration

Population
- → Finite population (countable units)
- → Infinite population. (uncountable units).
  ex - population of stars.

Population may be regarded as existent or hypothetical or imaginary population.
ex. The population of heads when a coin is tossed infinitely. Known as hypothetical.

* **Sample** :- part of population selected for representing the population in all its characteristics.

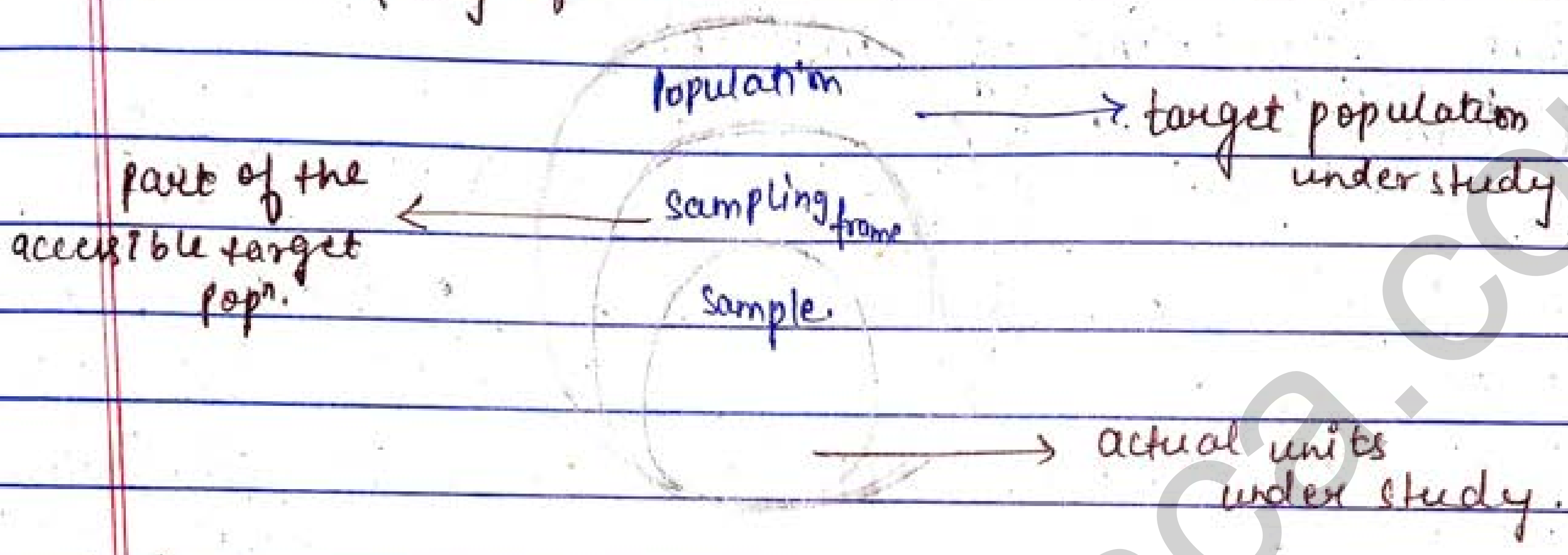
- Statistical inferences/decision about the population are drawn only on the basis of 'sample observation'.

- 'n' units $\xrightarrow{\text{represents}}$ 'Sample size.'

- Unit forming the sample → 'Sample units'

Sample unit :— jis cheez ki / jis cheez pr sampling ho rhi hai.

- A detailed and complete list of all the sampling unit Known as 'Sampling frame'.



* **Sampling** :- statistical procedure to infer/estimate about the unknown universe from a knowledge of a random sample drawn from it.

★ Principles of Sample Survey.

(1) Law of statistical regularity → reasonably large number of items selected at random from a large group of items will on the average represent the characteristics of the group/universe.

The reliability of a statistic in estimating a population characteristics varies as the square root of the sample size.

(2) Principle of Inertia → result drawn from the sample are likely more reliable, accurate as the sample size increases with other factors remain constant.

Page
Date

(3) Principle of optimization :- Optimum level of efficiency at min. cost or max. efficiency at a given level of cost achieved with appropriate sampling design.

(4) Principle of Validity :- Sampling design is valid only if it is possible to obtain valid estimates and valid tests about population parameters.
i.e. Random sampling.

★ Comparision b/w Sample Survey & Complete Enumeration

	Basis	Sample survey	Complete Enumeration.
(a)	Speed.	much more quickly because only part of vast population is enumerated.	Time-consuming because it involve all units.
(b)	Cost.	Individual unit cost → High (because better trained personnel are employed.) Total cost → low	Very-high.
(c)	Reliability	More reliable because of trained enumeratars, better supervision & application of modern technique.	Less reliable as compared to sample survey.
(d)	Necessity.	necessary when it comes to destructive sampling or sampling from a hypothetical popn.	Necessary to get detailed information & if the popn size is not large.

Page ______
Date ______

(c)	Accuracy.	Sampling error & Non-sampling error both occur in sample survey.	Sampling error is absent in complete enumeration. Only Non-sampling error occurs.

★ Errors in Sample Survey.

- Sampling Error
- Non-sampling error.

Sampling error → a statistical error that occurs when the analyst selects a sample that is not representative of the population being studied.

Factors affecting sampling error :-

(a) Error arising out due to defective sampling design i.e. bias or prejudice of the sampler.

(b) Error arising out due to substitution.

(c) Error owing to faulty demarcation of units.

(d) Error due to wrong choice of statistics.

(e) Variability in population.

Page
Date

Non-sampling errors

↓

Occurs in both sampling as well as complete enumeration.

Factors – kind of bias, Lapse of memory, ignorance, preference for certain digits, communication gap etc.

* Population $\xrightarrow{\text{characteristics}}$ Population Parameter (μ, S.D/σ^2, proportion)

↓ Random sampling.

Sample $\xrightarrow{\text{characteristics}}$ Sample statistic ($\hat{\mu}$, $\hat{S.D}$, $\hat{\sigma}^2$, $\hat{\text{proportion}}$) — estimate to popⁿ parameter

* **Parameter** :– A parameter is the characteristics of a popⁿ based on all the units of the popⁿ.

- Statistical inferences/decision are drawn about popⁿ parameter based on the sample observation drawn from the popⁿ.

- $$\boxed{\mu = \frac{\sum_{a=1}^{N} x_a}{N} = \frac{\text{sum of all values}}{\text{total population}}}$$

where, N = population size, μ = mean.

- $$\boxed{P = \frac{X}{N} = \frac{\text{Favourable no. of cases}}{\text{Total cases}}}$$

where, P = Proportion/Probability.

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

- $$\boxed{S.D(\sigma) = \sqrt{\frac{\sum (x_a - \mu)^2}{N}}}$$

Date

Note:- Avg → value corresponding to certain value of element.
Proportion → % or 0.0...
S.D or σ^2 → Kitna faila hai data.

* **Statistic** :- A statistic is used to estimate a particular population parameter.
The estimates of popⁿ mean, variance & popⁿ proportion are:-

- $\boxed{\bar{X} = \hat{\mu} = \frac{\Sigma x_i}{n}}$
- $\boxed{S_2 = \hat{\sigma}_2 = \frac{\Sigma (x_i - \bar{X})^2}{n}}$
- $\boxed{p = \hat{P} = \frac{x}{n} = \frac{\text{favourable no. of sample}}{\text{total samples}}}$

* Types of Sampling

1. Probability sampling. (Random sampling) ⇓ Simple Random Sampling (SRS)
2. Non-Probability sampling. (Non-Random sampling). ⇓ Purposive or Judgemental Sampling
3. Mixed sampling. ⇓ Systematic sampling.

→ Just/ Simple Random Sampling (SRS) - When units selected Independent of each other in such way that each units belonging to the popⁿ has equal chance of being part of sample.

SRS → SRS with replacement → one unit returned before the next unit is drawn so that popⁿ remain unchanged.

SRS → SRS without replacement → once the units selected from popⁿ one by one are never returned to the popⁿ before next drawing is made.

- SRS is very simple & effective method of drawing sample.

Condition of SRS -

(i) the popⁿ is not very large.
(ii) the sample size not very small.
(iii) popⁿ under consideration is not hetrogeneous.

→ Stratified sampling → dividing the popⁿ into number of Strata (groups) or sub-popⁿ in such a way that there should be very little variation among the units comprising a stratum & maximum variation should occur among diff. data.

Purpose -

(i) to make representation of all the sub-populations/strata.
(ii) to provide an estimate of parameter not only for all the strata but also and overall estimate.
(iii) reduction of variability & increase in precision.

Strata → groups.
Stratum → single group.

Date

Two types of allocation of sample size.

Bowley's Allocation
or
Proportional allocation.
↓
Variation in stratum is less.
↓
Sample size directly vary (∝) population size.

Neyman's allocation.
↓
Variation in stratum is more.
↓
Sample size ∝ popⁿ size & population S.D.

- Stratified sampling not advisable if –

(i) the popⁿ is not large.
(ii) some prior information not available.
(iii) Not much hetrogeneity among the units of popⁿ.

→ Multi Stage Sampling → sampling is carried out through stages.

Condition & features –
(i) Coverage is quite large.
(ii) It saves computational labour & ~~its~~ is cost-effective.
(iii) It adds flexibility into the sampling process which is lacking in other sampling schemes.
(iv) As compared to stratified sampling, it is likely less accurate.

Page
Date

Mixed/Systematic Sampling

↓

- unit constituting the sample are selected
  ↓
  at regular interval
  ↓
  after selecting the very first unit at random.

- Systematic sampling → partly probablity sampling (first unit is selected probabilistically)
  → partly non-probability sampling. (remaining unit are selected acc. to fixed rule which non-probabilistic in nature).

- if the popn size N is a multiple of sample size n.
  $\boxed{N = nK}$ where, K = sample interval.
  This type of systematic sampling known as "linear systematic sampling".

- if N is not multiple of n.
  $\boxed{N = nK + p}$ $p < K$.
  This type of systematic sampling known as "Circular systematic sampling".

Condition & Features-

(i) Very convinent when a complete & updated sampling frame is available.

(ii) less time consuming & less expensive.

Page
Date

Drawback. -

(i). if unknown & undetected periodicity in the sampling frame & sampling interval is a multiple of that period then we get biased sample.

(ii) it is not a probability sampling, no statistical inference / decision can be drawn about pop^n parameter.

* Sampling distribution and standard error of a statistic.

$$N \longrightarrow n$$

(population) (sample size)

- In case of sampling with replacement.

Total no. of samples $= \boxed{N^n}$

- In case of sampling without replacement.

Total no. of sampling $= \boxed{{}^N C_n}$

- The variation in the values of a statistic is termed as "sampling fluctuations".

- If it is possible to obtain the value of statistic (T) from all possible samples of a fixed sample size along with the corresponding probabilities, then statistic value to be treated as a random variable, in the form of a probability distribution.

Page
Date

- A probability distribution of statistic is known as sampling distribution of the statistic.

- The mean of the statistic
  ↓
  obtained from its sampling dist.
  ↓
  Known as "Expectation".

  The standard deviation of the statistic.
  ↓
  Known as the "Standard Error" (SE)

- SE regarded as measure of precision (extractness) achieved by sampling.
  S.E is inversely proportional to the square root of sample size. , Sample size ↑ SE ↓

$$\boxed{SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}} \text{ for SRS WR}$$

$$\boxed{SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \text{ for SRS WOR}$$

- Standard Error for Proportion

$$\boxed{S.E(p) = \sqrt{\frac{pq}{n}}} \text{ for SRS WR}$$

$$\boxed{S.E(p) = \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}} \text{ for SRSWOR}$$

Factor $\sqrt{\frac{N-n}{N-1}}$
↓
Known as
↓
Finite population correction (fpc)
or
Finite population multiplier.

Date

Measures of Central Tendency and Dispersion

Mean

Arithmetic mean | Harmonic mean | Geometric mean.

* Arithmetic mean :-

$$AM = \frac{\text{Sum of all observations}}{\text{Total no. of observations}} = \frac{\Sigma x_i}{N(\Sigma f)}.$$

$$\boxed{\bar{x} = \frac{\Sigma x_i}{N}} \quad (i=1)$$

Arithmetic mean of frequency distribution

$$\boxed{\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}} \quad [\text{Direct method}].$$

$$\boxed{\bar{x} = A + \frac{\Sigma d_i f_i}{\Sigma f_i} \times h} \quad (\text{step-deviation method}).$$

where, A = assumed mean

$$d_i = \frac{x_i - A}{h}$$

Properties of AM.

(i) If the observation are say 'k' then AM is also 'k'.

(ii) The sum of deviation (diff) from the mean is zero

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

(iii) If all the observations are added (+), subtracted (-), or multiplied by value K, then AM also gets added, subtracted or multiplied by value 'K'.

(iv) AM is affected due to change in origin and/or scale which implies that if the origin variable x is changed to another variable y by effecting a change in origin say a, and scale b of x i.e. $y = a + bx$ then A.M of y is given by

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

origin shift ← $y = a + bx$ → scaling

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

(v) If there are 2 groups containing n_1 & n_2 observations and $\bar{x}_1$ and $\bar{x}_2$ as the respective A.M then combined AM is -

$$\boxed{\bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}}$$

★ Correcting Incorrect mean

$$\boxed{\bar{x}_{r} = \bar{x}_{w} + \frac{x_{r} - x_{w}}{n}}$$

★ weighted AM. $\boxed{\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i}}$

Date

Arithmetic mean

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, AM is affected least by fluctuations of sampling.

Demerits

1. Affected by extreme values
2. Open-end classes
3. Not detected graphically
4. Qualitative data (not computed).
5. Lead to wrong conclusions if details are not available.

$\boxed{\text{MEDIAN}}$ → middle value, positional value.

* Ungrouped data (data in ascending order)

n is odd, $\left(\frac{n+1}{2}\right)^{th}$ term

n is even, $\dfrac{\frac{n}{2}^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$ term

$\left(\frac{n+1}{2}\right)^{th} \Rightarrow 2.5^{th} = 2^{nd} + 0.5\,(3^{rd} - 2^{nd})$.

* Discrete Series

$cf \geq \frac{N+1}{2}$ → median value will be x corresponding to it.

Date

* Continuous Series

(i) Find cumulative frequency

(ii) $\left(\frac{N}{2}\right)^{th}$ value

(iii) $cf \geq \frac{N}{2}$ → median class

(iv) $$\boxed{\text{median} = l + \left(\frac{\frac{N}{2} - cf_{(c-1)}}{f_1}\right) \times h}$$

$cf_{(c-1)}$ → CF of class before.

where, l = lower class boundary of median class

$h = l_2 - l_1$ (class-length)

Properties of median

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

→ $\boxed{y_{median} = a + bx_{median}}$

(ii) For a set of observations, the sum of absolute (sign ignore) deviation is minimum when the deviation are taken from the median.

$\Sigma(x - x_{median})$ → minimum.

Median

Merits

- rigidly defined
- simple to calculate
- Can deal with open-end class.
- unaffected by extreme value
- Qualitative data (Ranking)
- Determined graphically.

Demerits

- Not suitable for further mathematical treatment
- Not based on each & every item
- much affected by sampling fluctuations.
- need to arrange data.
- Ungrouped data — even not suitable.

Date

Partition Value

(i) Quartiles (3 quartiles) → divide in 4 equal part (25% each)

(ii) Deciles (9 Deciles) → divide in 10 equal part (10% each)

(iii) Percentiles (99 percentiles) → 100 equal parts (1% each).

- Calculation of Quartiles (individual observations)

$$K^{th} \text{ quartile} = K\left(\frac{n+1}{4}\right)^{th} \text{ value.}$$

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

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

- Calculation of Deciles:-

$$\boxed{D_K = \left[K\left(\frac{n+1}{10}\right)^{th}\right]} \qquad \underline{K = 1 \text{ to } 9}$$

- Calculation of Percentiles:- $\boxed{P_K = K\left(\frac{n+1}{100}\right)^{th}}$ $\underline{K = 1 \text{ to } 99}$

* In case of Discrete Series:-

→ Quartiles Q_K → value corresponding to $CF \geq \left(\frac{K(N+1)}{4}\right)$

→ Deciles D_K → value corresponding to $CF \geq K\left(\frac{N+1}{10}\right)$

→ Percentiles P_K → value corresponding to $CF \geq K\left(\frac{N+1}{100}\right)$

Date

* Continuous Series

$$\boxed{x_{mode} = l_1 + \frac{f_0 - f_{-1}}{2f_0 - f_1 - f_{-1}} \times h}$$

Limitations of this for :- (i) Exclusive type series
(ii) Same length of class interval.

Properties

(i) Mode + Mean = 3(mean - median)
mean - Mode = 3(mean - median)
OR
3 Median = 2 Mean + Mode

Empirical formula

moderately skewed Distribution

Symmetrical

mode < median < mean

Positively skewed.

mean < median < mode

Negatively skewed.

(ii) $y = a + bx$, then $\boxed{y_{mode} = a + bx_{mod}}$

★ In case of Continuous Series

- Quartiles

$$x_{median}\,(Q_K) = l + \frac{\frac{KN}{4} - cf_{-1}}{f} \times h$$

$cf \geq \frac{KN}{4}$

where, l = lower limit of median class
cf = c.f. of class proceeding the median class.
h = ~~fxxx~~ class length.
f = frequency of median class.

- Deciles.

$$x_{median}\,(D_K) = l + \frac{\frac{KN}{10} - cf_{-1}}{f_1} \times h$$

$cf \geq \frac{KN}{10}$

- Percentile

$$x_{median}\,(P_K) = l + \frac{\frac{KN}{100} - cf_{-1}}{f_1} \times h$$

$cf \geq \frac{KN}{100}$

Imp: Ogives help us to find Quartiles graphically.

MODE → most occuring value.

* Bi-model distribution – 2 modes
multi model " – more than 2 modes
No mode – ex→ 1, 2, 8, 6, 3. (1 observation of each).

★ Discrete Series
↓
max frequency.

* **Merits of mode:-**

- Simple and easy to calculate.
- Locate by inspection
- Graphically by histogram.
- Not affected by extreme value.
- Open - end class not affected.

* **Demerits of mode:-**

- Not rigedly defined
- Not based on all observations
- Not for further mathematical treatment.
- Affected by sampling.
- May have bi-modal distribution.

*Imp. Most popular - Arithmetic mean
Most reasonable - Geometric mean.

Geometric Mean

- $GM = (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{\frac{1}{n}}$.
  where n = no. of observations
- $\log GM = \frac{1}{n} \Sigma \log x_i$

  $GM = \text{antilog}\left(\frac{\Sigma \log x_i}{n}\right)$

Individual series.

- $GM = (x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdots x_n^{f_n})^{\frac{1}{n}}$

  $GM = \text{antilog}\left(\frac{\Sigma f_i \log x_i}{N}\right)$

Discrete series.

- Continuous Series = same except x_i = mid-point in continuous series.

Page
Date

Uses:-
(i) When ratio of consecutive terms remains constant.
(ii) most appropriate average in index number.
(iii) Give weightage to smaller items.

Properties of GM

(i) If observations are 'K' then GM is also 'K'.
(ii) If $z = xy$, then $GM_z = GM_x \times GM_y$.
(iii) If $z = \frac{x}{y}$ then $GM_z = \frac{GM_x}{GM_y}$.

Note- If any observation is zero then G.M is also zero

Harmonic Mean

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} = \boxed{\frac{n}{\Sigma \frac{1}{x_i}}}$$

★ Discrete Series $\Rightarrow$ $HM = \frac{\Sigma f_i}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots \frac{f_n}{x_n}} = \frac{\Sigma f_i}{\Sigma \frac{f_i}{x_i}}$

★ Continuous Series $\Rightarrow$ $HM = \frac{\Sigma f_i}{\Sigma \frac{f_i}{x_i}}$ (where x_i = mid-point)

Uses:-
When ratio of the quantities are given in which numerator is constant.

Date

Properties of HM.

(i) If observation are 'K' then HM is also 'K'.

(ii) n_1 n_2

HM_1 HM_2 $HM_{combined} = \dfrac{n_1+n_2}{\dfrac{n_1}{H_1}+\dfrac{n_2}{H_2}}$

Weighted Mean

$\boxed{AM_w = \dfrac{\Sigma w_i x_i}{\Sigma w_i}}$, $\boxed{GM_w = \text{antilog}\dfrac{\Sigma w_i \log x_i}{\Sigma w_i}}$

$\boxed{HM_w = \dfrac{\Sigma w_i}{\Sigma \dfrac{w_i}{x_i}}}$

★ Relationship between AM, GM and HM.

(1) if all observations are equal 'K' then, GM = HM = AM.
else, AM ≥ GM ≥ HM

(2) For two numbers, $\boxed{GM^2 = AM \cdot HM}$

Note:- if any data -ve then, GM will not work because answer will not be real number.

if any value in a given data is 0 then HM not defined

Page
Date

★ Sum of 1^{st} n natural no = $n\frac{(n+1)}{2}$

$$1+2+3+\dots+n = n\frac{(n+1)}{2}$$

★ Sum of square of 1^{st} n natural number $= \frac{n(n+1)(2n+1)}{6}$

$$1^2+2^2+3^2+4^2\dots\dots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

★ Sum of cubes of 1^{st} n natural number $= \left(\frac{n(n+1)}{2}\right)^2$

$$1^3+2^3+3^3+4^3\dots\dots+n^3 = \left(\frac{n(n+1)}{2}\right)^2$$

Ques:- If a variable assumes the values 1, 2, 3 ... 5 with weights as 1, 2, 3 ... 5, then what is the AM ?

X	Wi	XiWi
1	1	1^2
2	2	2^2
3	3	3^2
4	4	4^2
5	5	5^2

$$AM = \frac{\Sigma W_i x_i}{\Sigma W_i} \Rightarrow \frac{1^2+2^2+3^2+4^2+5^2}{1+2+3+4+5}$$

$$\Rightarrow \frac{\frac{n(n+1)(2n+1)}{6}}{\frac{n(n+1)}{2}} \Rightarrow \frac{\frac{5(5+1)(2\times5+1)}{6}}{\frac{5(5+1)}{2}}$$

$$\Rightarrow \frac{\cancel{5}\cancel{(5+1)}(10+1)}{\cancel{6}\,3} \times \frac{\cancel{2}\,1}{\cancel{5}\cancel{(5+1)}} \Rightarrow \boxed{\frac{11}{3}}$$

Dispersion (scatterness)

↓

amount of scatterness of the observation.

Measures of Dispersion.

Absolute ↓

- dependent on the unit of the variable under consideration.
- Easy to comprehend and compute.
- Diff measures ways :-
  - → Range
  - → mean deviation
  - → Standard deviation
  - → Quartile deviation

Relative ↓

- Unit free
- For comparing two or more distributions, relative measures of dispersion considered.
- Diff measures ways :-
  - → Coefficient of Range
  - → " of mean deviation
  - → Coefficient of Variation
  - → Coefficient of Quartile deviation

RANGE

Range = Highest class boundary − Smallest class boundary.

$$\boxed{\text{Range} = \text{Largest} - \text{Smallest}}$$

$$\text{Coefficient of Range} = \boxed{\frac{L-S}{L+S} \times 100}$$

$L \to H.C.B$

$S \to S.C.B$

Properties of Range-

$y = a + bx$

$\boxed{\text{Range of } y = |b| \text{ Range of } x}$, $\boxed{R_y = |b| R_x}$

Range

Merits

- Simple to understand
- Easy to calculate
- minimum time to calculate

Demerits

- Not based on all observation.
- Considers only the extreme values.
- affected by fluctuation of sampling
- Cannot calculated with open-end classes.
- Not suitable for further mathematical treatment.

MEAN DEVIATION

★ Individual series.

$$\boxed{MD_{\bar{x}} = \frac{\sum |x_i - \bar{x}|}{n}} \qquad \boxed{MD_{median} = \frac{\sum |x_i - M_d|}{n}}$$

★ Discrete series.

$$\boxed{MD_{\bar{x}} = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i}} \qquad \boxed{MD_{med} = \frac{\sum f_i |x_i - x_{median}|}{\sum f_i}}$$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \qquad x_{median} = cf \geq \frac{n+1}{2}$$

$n = \sum f$.

★ Continuous series

$$\boxed{M \cdot D \text{ (about an average } A) = \frac{\sum f_i |x_i - A|}{\sum f_i}}$$

★ Coefficient of mean deviation

$$\boxed{\text{coefficient of mean/median} = \frac{m.D}{\bar{x}/\text{median}} \times 100}$$

★ Properties of m.D.

- Mean deviation takes its minimum value when the deviation are taken from the median.
- $y = a + bx$, a and b belog constant, then.

$$\boxed{\text{mD of } y = |b| \text{ mD of } x}$$

★ Merits of m.D
- easy and simple - rigidly defined
- based on all observations.
- less affected by extreme deviation.

Demerits of m.D.
- algebric signs are ignored while taking deviation.
- not suitable for further mathematical treatment
- cannot be calculated for open-end class.

Imp. Best method - mean deviation from median.

QUARTILE DEVIATION OR SEMI INTER-QUARTILE

Inter-quartile range = $Q_3 - Q_1$

Quartile deviation / Semi inter quartile = $\boxed{\frac{Q_3 - Q_1}{2}}$

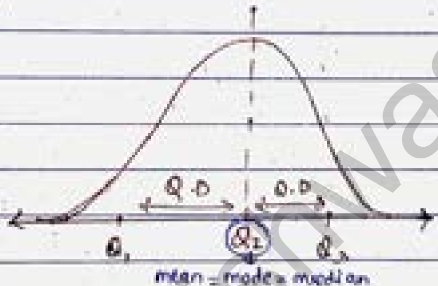
Date

* **Coefficient of Q.D.**

$$\boxed{\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100}$$

Coefficient of Q.D is a pure number independent of the units of measurement and can be used to compare two distribution expressed in different.

* **Symmetrical distribution**



* $\boxed{Q.D = Q_2 - Q_1}$
* $\boxed{Q.D = Q_3 - Q_2}$

* $QD - QD = Q_2 - Q_1 - Q_3 + Q_2$

  $Q_1 + Q_3 = 2Q_2$

* $\boxed{\underset{\text{(median)}}{Q_2} = \frac{Q_1 + Q_3}{2}}$

* $\boxed{QD = \frac{Q_3 - Q_1}{2}}$

Coefficient of Q.D.

$$COQD = \frac{Q_3 - Q_1/2}{Q_3 + Q_1/2} \times 100$$

* $\boxed{COQD = \frac{Q.D}{Q_2} \times 100}$

or

* $\boxed{C.OQD = \frac{Q.D}{\text{median}} \times 100}$

★ Properties of Q.D

- $y = a + bx$ — a: change of origin; bx: change of scale.

  Q.D remains unaffected due to a change of origin but is affected in the same ratio due to change in scale.

  $$\boxed{Q.D_y = |b| \, QD_x}$$

Quartile Deviation.

Merits

- Best measure of dispersion for open-end classification
- Less affected by sampling fluctuations.
- Useful when it is desired to study variability in the central half part of the data.

Demerits

- Not based on all observations.
- Not suitable for further mathematical treatment
- Considerably affected by sampling fluctuations.

STANDARD DEVIATION

Standard Deviation → Root mean squared deviation from mean.

Denoted by sigma (σ)

$$\boxed{S.D\ (\sigma) = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}} \quad \text{or} \quad \boxed{S.D\ (\sigma) = \sqrt{\frac{\Sigma x_i^2}{n} - (\bar{x})^2}}$$

In case of Frequency distribution -

$$\boxed{\sigma = \sqrt{\frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i}}} \quad \text{or} \quad \boxed{\sigma = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2}}$$

Page
Date

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

calculator trick :- $f_i x_i^2 \Rightarrow \boxed{x_i \times = \times f_i = M+}$

* Continuous Series
Same Formula except x_i midvalue

Variance

The variance of a given set of observation is defined as the square of its S.D.

Denoted by $\boxed{\sigma^2}$

Individual observation $\sigma^2 = \boxed{\frac{\Sigma (x_i - \bar{x})^2}{n}}$ or $\boxed{\frac{\Sigma x_i^2}{n} - (\bar{x})^2}$

Ungrouped / Grouped $\sigma^2 = \boxed{\frac{\Sigma f_i x_i^2}{n} - \left(\frac{\Sigma f_i x_i}{n}\right)^2}$ or $\boxed{\frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i}}$

or

$\boxed{\frac{\Sigma f_i x_i^2}{n} - (\bar{x})^2}$ where $n = \Sigma f_i$

Imp. points :- For any two numbers S.D is always Half of the range.

Ques. The S.D of 10, 16, 10, 16, 10, 10, 16, 16

x	f
10	4
16	4

$S.D = \frac{|a-b|}{2} \Rightarrow \frac{|10-16|}{2} \Rightarrow \frac{|-6|}{2} \Rightarrow \frac{6}{2} = \boxed{3}$

(If frequency of any data is same we ignore frequency)

Page:
Date:

★ GRW

Imp Ques. The mean and S.D for a, b and 2 are 3 and $\frac{2}{\sqrt{3}}$ respe
The value of ab would be.

Given

$\Rightarrow \bar{x} = \frac{a+b+2}{3} = 3 \Rightarrow a+b+2 = 9.$
$a+b = 9-2 = 7. \Rightarrow \boxed{a+b=7}$

$\Rightarrow S.D = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2} \Rightarrow \left(\frac{2}{\sqrt{3}}\right)^2 = \left(\sqrt{\frac{a^2+b^2+2^2}{3} - (3)^2}\right)^2$

$\Rightarrow \frac{2^2}{(\sqrt{3})^2} = \frac{a^2+b^2+4}{3} - 9$

$\Rightarrow \frac{4}{3} = \frac{a^2+b^2+4-27}{3} \Rightarrow \frac{4}{\cancel{3}} \times \cancel{3} = a^2+b^2+4-27.$

$\Rightarrow 4-4 = a^2+b^2-27 \Rightarrow \boxed{a^2+b^2=27} \rightarrow eq.2$

$(a+b)^2 = (7)^2 \rightarrow$ from mean eq.①
$a^2+b^2+2ab = 49$ (put value from eq. 2)
$27+2ab = 49.$
$2ab = 49-27$
$2ab = 22$
$ab = \frac{22}{2} = \boxed{11}$ Ans

Page

Date

★ Correcting Incorrect Values of Mean and Standard Deviation

Given:- $\sigma_w, \bar{x}_w, n, x_w, x_r$.

Steps:- 1) $\bar{x}_r = \bar{x}_w + \frac{x_r - x_w}{n}$

2) $\sigma_w = \sqrt{\frac{\Sigma x_w^2}{n} - (\bar{x}_w)^2}$ → $\sigma_w, n, \bar{x}_w$ (given)
Find Σx_w^2

3) $\Sigma x_r^2 = \Sigma x_w^2 - x_w^2 + x_r^2$

4) $\sigma_r = \sqrt{\frac{\Sigma x_r^2}{n} - (\bar{x}_r)^2}$

★ Combined Standard Deviation

$$\boxed{\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}}$$

where, $d_1 = \bar{x}_1 - \bar{x}$
$d_2 = \bar{x}_2 - \bar{x}$

★ Coefficient of Variation

$$\boxed{\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100}$$

This is a pure number independent of units of measurement and hence can be used to compare the variability of two distribution expressed in diff uses.

COV ↑ Less consistent / Less uniform

⇒ Coefficient of variance → smaller → Less variable → More consistent

⇒ Coefficient of variance → greater → more variable → Less consistent

Imp. Que. The algebraic sum of 10 items about deviation 8 is −5. Find its A.M. Also COV if SD is 1.5.

⇒ $n = 10$

$\Sigma(x_i - 8) = -5$

$\Sigma x_i - 8 \times n = -5$ $\quad (\Sigma(x_i - \bar{x}) = 0 \Rightarrow \Sigma x_i - n\bar{x} = 0)$

$\Sigma x_i - 8 \times 10 = -5$

$\Sigma x_i = 80 - 5$

$\Sigma x_i = 75$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{75}{10} = \boxed{7.5}$

$\sigma = 1.5,\ \bar{x} = 7.5$

$COV = \frac{\sigma}{\bar{x}} \times 100 = \frac{1.5}{7.5} \times 100 = \boxed{20}$ Ans

Properties of standard deviation

- Suitable for further mathematical treatment

- $y = a + bx$

  $\boxed{SD_y = |b|\, SD_x}$

- If all observation assumed by a variable are equal, then S.D is zero.

  (This property applies in all MOD).

Page______
Date______

- The S.D. of first n natural number is –

$$\boxed{S.D = \sqrt{\frac{n^2-1}{12}}}$$

★ Relationship between S.D, MD, Q.D :

$\boxed{4SD = 5MD = 6Q.D}$ (applicable only in symmetrical distribution).

Imp.- If all the observation is increase / decrease by any number, then (SD, MD, Q.D) remain unchanged.

Probability.

The terms 'probably', 'chance', 'odd in favour', 'odd against' are too familiar nowadays and they have their origin in a branch of Mathematics known as probability.

★ Event → Result

★ Experiment → a process with result.

★ Random experiment → a process in which result is not known in advance.

★ Events :-

- **Simple**
  → only 1 ans
  ex. H | T
- **Composite**
  → multiple ans.
  ex. 1H & 1T

★ Mutually exclusive Events or Incompatible events :-

→ Kuch common nhi hai.

→ Order matter krta hai. $\begin{pmatrix} H & T \\ T & H \end{pmatrix}$ both are different.

★ Exhaustive events :-

→ all possible events of A

e.g. coin toss → H, T   Dice throw → 1, 2, 3, 4, 5, 6.

Page
Date

★ Equally likely events (unbaised event)
→ hone ki possibility बराबर हो।

✳ <u>Classical defination of Probability.</u>

- For finite elementary events:-
  - finite → countable
  - elementary → simple events

$$\boxed{P(A) = \frac{n_A}{n}} = \frac{\text{no. of equally likely events favourable to A}}{\text{total no. of equally likely events.}}$$

- For finite composite events:-

$$\boxed{P(A) = \frac{m_A}{m}} = \frac{\text{no. of mutually exclusive, exhaustive \& equally likely event favourable to A.}}{\text{total no. of mutually exclu., exhaustive \& equally likely events}}$$

e.g. 2 coin toss

①	②
H	H
H	T
T	H
T	T

Ⓐ → Both heads.

Finite simple events $\boxed{P(A) = \frac{n_A}{n} = \frac{1}{4}.}$

Ⓑ → 1H + 1T

For finite composite events. $\boxed{P(B) = \frac{2}{4}}$

Page
Date

Broad division of probability.

- Subjective probability
- Objective probability.

Imp. note:- (a) $\boxed{0 \leq P(A) \leq 1}$ probability lies b/w 0 to 1.

$P(A) = 0$ impossible events in A.

$P(B) = 1$ sure event in B.

(b) $\boxed{P(A) + P(A') = 1}$ $P(A)$ = happening of event

$P(A')$ = non-happening of event.

(c) Odd in favour = $\boxed{n_A : n_B}$

Odd in favour = $\dfrac{\text{no. of favourable events}}{\text{no. of unfavourable events}} = \dfrac{n_A}{n_B}$.

↓ Inverse.

odd in against = $\boxed{\dfrac{n_B}{n_A}}$

A→odd in favour $\boxed{P(A) = \dfrac{n_A}{n_A + n_B}}$

B→odd in against $\boxed{P(B) = \dfrac{n_B}{n_A + n_B}}$

* **Demerits.**

(i) applicable only when total no. of events is finite.

(ii) used only when the events are equally likely or equi-probable.

(iii) limited → where the possible events are known in advance.

e.g. coin tossing, dice throwing, etc.

Page
Date

* Set - Theoretic approach to Probability.

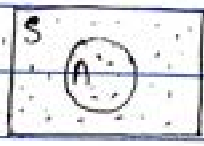
$$\boxed{P(A) = \frac{n(A)}{n(S)}}$$

where,

$n(A)$ = no. of count favourable to A.

universal set ↓ sample space. Ω or S → sample space.

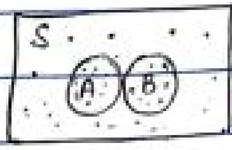
$n(S)$ = total no. of count in space.



- **Two events A and B are mutually exclusive.**

$$\boxed{P(A\cup B) = \frac{n(A\cup B)}{n(S)}} = \frac{n(A)+n(B)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)}$$

$$\boxed{P(A\cup B) = P(A) + P(B)}$$

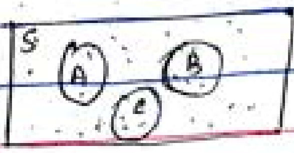


$$P(A\cap B) = \frac{n(A\cap B)}{n(S)} = \frac{0}{n(S)} = 0.$$

- **Three events A, B and C are mutually exclusive.**

$$P(A\cup B\cup C) = \frac{n(A\cup B\cup C)}{n(S)} = \frac{n(A)+n(B)+n(C)}{n(S)} \Rightarrow \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} + \frac{n(C)}{n(S)}$$

$$\boxed{P(A\cup B\cup C) = P(A) + P(B) + P(C)}$$



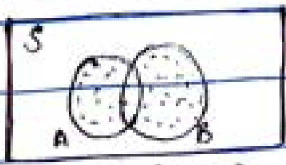
$$\boxed{P(A\cap B\cap C) = 0}$$

Page
Date

Exhaustive
↓
A & B → all elements of are in A or B or $A \cap B$.

- Two events A and B are exhaustive.

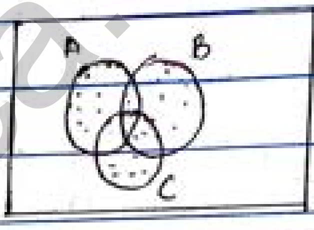
$$\boxed{P(A \cup B) = \frac{n(A \cup B)}{n(S)} = 1}$$



$n(A \cup B) = n(S)$

- Three events A, B & C are exhaustive.

$$\boxed{P(A \cup B \cup C) = \frac{n(A \cup B \cup C)}{n(S)} = 1}$$



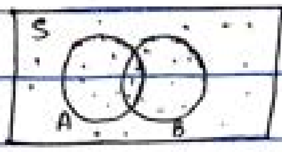
$n(A \cup B \cup C) = n(S)$.

- Three events are equally likely.

$$\boxed{P(A) = P(B) = P(C)}$$

- if A & B are not mutually exclusive → $n(A \cap B) \neq 0$.

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$



- If A, B, C are not mutually exclusive & neither they are subset

$$P(A \cup B \cup C) = \frac{n(A \cup B \cup C)}{n(S)}$$

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)}$$

Page
Date

* Axiomatic or modern definition of Probability.

(i) P(A) lies b/w 0 and 1 for every ⊂ subset

$$\boxed{\begin{array}{l} P(A)=0 \\ \rightarrow n(A)=0 \end{array}} \qquad \boxed{\begin{array}{l} P(A)=1 \\ \rightarrow n(A)=n(S) \end{array}}$$

(ii) $P(S) = \frac{n(S)}{n(S)} = 1$ $\quad \boxed{P(S)=1}$

(iii) $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$

- **Theorem - 1** (for any two mutually exclusive events)

$$\boxed{P(A \cup B) \text{ or } P(A+B) = P(A) + P(B)}$$

because $P(A \cap B)$ or $P(A \cdot B) = 0$.

Note:- if in b/w two events → $+ \cup$ → 'or' या
$\times \cap$ → 'and' और

- **Theorem - 2** (mutually exclusive events).

$$\boxed{P(A_1 \cup A_2 \cup \ldots A_K) = P(A_1) + P(A_2) + P(A_3) + \ldots P(A_K)}$$

- **Theorem - 3** (for any two event)

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

- **Theorem - 4.** (for any 3 event)

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)}$$

* Compound Probability or Joint Probability.

- Dependent event → if two or more events are there & performing a event is depend on the other.

- Independent event → If two or more events are happening & there is no dependency of event on each other.

$P(A)$ = independent $\quad\quad$ $P(B/A)$ B is dependent of A.

* Dependent event (conditional probability)

$P(B-A) = P(B) - P(A\cap B)$

$$\boxed{P(B/A) = \frac{P(A\cap B)}{P(A)}}$$

$$\boxed{P(A/B) = \frac{P(A\cap B)}{P(B)}}$$

$$P(A'/B) = \frac{P(A'\cap B)}{P(B)} = \frac{P(B-A)}{P(B)}$$

$$P(A/B') = \frac{P(A\cap B')}{P(B')} = \frac{P(A-B)}{P(B')}$$

$$P(A'/B') = \frac{P(A'\cap B')}{P(B')} = \frac{P(S)-P(A\cup B)}{P(B')}$$

Demorgan's Law.

$$\boxed{\begin{aligned} P(A'\cap B') &= P(A\cup B)' \\ P(A'\cup B') &= P(A\cap B)' \end{aligned}}$$

$$P(B/A) = \frac{P(A\cap B)}{P(A)} \implies P(A\cap B) = P(A)\cdot P(B/A).$$

(if B is depending on A).

$$P(A/B) = \frac{P(A\cap B)}{P(B)} \implies P(A\cap B) = P(B)\cdot P(A/B).$$

(if A is depending on B).

$P(A-B) = P(A) - P(A\cap B)$
$P(B-A) = P(B) - P(A\cap B)$

Page
Date

★ Independent Event

$P(B/A) = P(B)$

$P(A/B) = P(A)$

$P(A\cap B) = P(A)\cdot P(B)$

$P(A \text{ and } B') / P(A\cap B') = P(A)\cdot P(B') = P(A)\cdot(1-P(B))$

$P(A' \text{ and } B) / P(A'\cap B) = P(A')\cdot P(B) = (1-P(A))\cdot P(B)$

$P(A' \text{ and } B') / P(A'\cap B') = P(A')\cdot(B')$

Note :- if question is silent assume independent event.

Independent → $P(A\cap B\cap C) = P(A)\cdot P(B)\cdot P(C)$

Compound Probability of 3 events

Dependent

$\boxed{P(A\cap B\cap C) = P(A)\cdot P(B/A)\cdot P(C/B\cap A)}$

here, A is independent, B is dependent on A, C is dependent on B & A

$P(B/A) = \frac{P(A\cap B)}{P(A)}$ $\quad$ $P(C/B\cap A) = \frac{P(C\cap B\cap A)}{P(B\cap A)}$

$P(A'\cap B') = P(A\cup B)'$

$\Rightarrow \boxed{P(A\cup B) = P(S) - P(A'\cap B') = 1 - P(A'\cap B')}$

↓

$P(A\cup B) = 1 - P(A\cup B)'$

$\Rightarrow \boxed{P(A\cup B) + P(A\cup B)' = 1}$

$\Rightarrow \boxed{P(A\cup B\cup C) = P(S) - P(A'\cap B'\cap C')}$

* Random Variable / Stochastic variable.

Experiment → samplespace → function → Random variable / ~~Ans (Real no)~~.

e.g. 3 times coin toss

X → no. of heads in an event.

HHH	X = 3
HHT	X = 2
HTH	X = 2
HTT	X = 1
TTT	X = 0
THT	X = 1
TTH	X = 1
THH	X = 2.

X	0	1	2	3
P(x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
	P_1	P_2	P_3	P_4.

$$P_1 + P_2 + P_3 + P_4 = \sum_{i=1}^{4} P_i$$

$$\Rightarrow \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} \Rightarrow \frac{8}{8} = 1$$

Imp:- (i) $\boxed{0 \le P_i \le 1}$

(ii) $\boxed{\sum P_i = 1}$

Random variable.

Discrete Random variable	Continuous Random Variable.
(fixed value. e.g. no. of accident etc)	(interval e.g. height, weight etc).
* $X = x_1, x_2, x_3 \cdots x_n$.	* $X = [\alpha, \beta]$
* $P(X) = P_1, P_2, P_3 \cdots P_n$. $P_1 = P(x_1)$, $P_2 = P(x_2)$, $P_3 = P(x_3)$.	* Probability density function (p.d.f) = $f(x)$.
* $P_i \ge 0$	* $f(x) \ge 0$
* $\sum P_i = 1$	* $\int_{\alpha}^{\beta} f(x)\,dx = 1$

* Probability Density function (PDF)

(i) $f(x) \geq 0$ for $x \in [\alpha, \beta]$

(ii) $\int_{\alpha}^{\beta} f(x)\,dx = 1$

$\int$ (integration sign)

* Expected Value

(i) $P_i \geq 0$ for every x_i

(ii) $\Sigma P_i = 1$

$\boxed{\mu \text{ or } E(x) = \Sigma p_i x_i}$ $\left(\mu = \dfrac{P_1 x_1 + P_2 x_2 + P_3 x_3 - \cdots + P_n x_n}{P_1 + P_2 + P_3 - \cdots + P_n}\right) = \dfrac{\Sigma P_i x_i}{\Sigma P_i}$

mean $(\Sigma P_i = 1)$

$\boxed{E(x^2) = \Sigma p_i x_i^2}$

Variance of 'x' → $(V(x))$ or σ^2

$$\boxed{\begin{array}{l} V(x) \text{ or } \sigma^2 = E(x^2) - (E(x))^2 \\ \text{where, } E(x^2) = \Sigma p_i x_i^2 \\ \quad \& \ E(x) = \Sigma p_i x_i \end{array}}$$

positive square root of variance is S.D

$\boxed{\sigma = \sqrt{V(x)} = \sqrt{E(x^2) - (E(x))^2}}$

$E[g(x)] = \Sigma(p_i(g(x_i)))$

Properties :-

- if $y = a + bx$
  $\mu_y = a + b\mu_x$
- if $y = a + bx$
  $\sigma_y = |b|\,\sigma_x$
  $\sigma_y^2 = b^2 \sigma_x^2 \quad (\sigma_y^2 = V(y))$

* Probability mass function.

$p_i = f(x) =$ probability mass function.

$\mu = E(x) = \sum p_i x_i = \sum x(f x)$

$E(x^2) = \sum p_i x_i^2 = \sum x_i^2 (f(x_i))$

$\sigma^2 = E(x^2) - (E(x))^2$
$= E x_i^2 (f(x_i)) - (\sum x_i f(x_i)^2)$

$\sigma^2 = \sum p_i (x_i - \mu)^2$
$= \sum p_i (x - (\sum p_i x_i))^2$

if x is continuous random variable.

Dis.		continuous.
$\sum$	$\longrightarrow$	$\int (\ ) dx$
p_i	$\longrightarrow$	$f(x)$

$\mu = E(x) = \int x f(x)\, dx$.

$E(x^2) = \int x^2 f(x)\, dx$.

$\sigma^2 = \int x^2 f(x)\, dx - \left(\int x f(x) dx\right)^2$

Imp. points → 1. Expectation of all values equal to constant K.

$X \rightarrow K, K, K \ldots$

$E(K) = K \qquad \sigma^2 = 0, \ \sigma = 0$

2. $E(x+y) = E(x) + E(y)$
3. $E(Kx) = K(E(x)) \longrightarrow$ e.g. $E(3x) = 3(E(x))$
4. $E(xy) = E(x) \cdot E(y)$

Theoritical Distribution

Uses –
- Life span
- future projection
- Decision making on some sample → Population.

Discrete Random variable → * Binomial distribution, * Poisson Distribution

Continuous Random Variable → * Normal distribution

* Binomial distribution.

↓

derived from a particular type of random experiment called Bernoulli process named after famous mathematician Bernoulli.

When 'trial' is attempted to produce a particular outcome which is neither certain nor impossible.

Trial / Random experiment
↓
Outcome / event
→ success p
→ failure $q = 1-p$.

Formula. → $\boxed{P(x) \text{ or } f(x) = {}^nC_x\, p^x q^{n-x}}$ where, n = no. of random experiment, x = no. of success.

p = success probability of an event.
q = failure probability of an event.

Imp. note →

- $p+q=1$
  - $0<p<1$
  - $0<q<1$
- $q = 1-p$.
- if $n>0$, $p, q>0 \longrightarrow f(x) \geq 0$.
- $\Sigma f(x) = 1$.
- Binomial distribution also known as $\boxed{\text{bi-parametric}}$ distribution as it is characterised by two parameters: $\boxed{n \,\&\, p.}$

if in Discrete Random variable $P(x)$ written as $f(x)$
$\downarrow$
∴ 'Probability mass function'

- mean of binomial dist. $\boxed{(\mu) = np}$ or $\boxed{E(x) = np}$
- Variance of binomial dist. $\boxed{(\sigma^2) = npq}$
- Standard deviation $\boxed{(\sigma) = \sqrt{npq}}$
- Variance of a binomial variable is less than its mean.

$$\boxed{\sigma^2 < \mu}$$

- Mode → highest freq value of x.

* if $(n+1)p$ is a non-integer

Mode = greatest integer of $(n+1)p$. ex → $(n+1)p = 7.8$

mode = 7 Ans.

* if $(n+1)p$ is an integer.
  Mode $= (n+1)p$ or $(n+1)p - 1$ $\longrightarrow$ Bimodal distribution

ex $\rightarrow$ $(n+1)p = 8$.
mode $=$ 8 or 7 Ans.

- $X \sim B(n, p) \longrightarrow P(x) = {}^{n}C_x \, p^x q^{n-x}$

&

$Y \sim B(n_2, p) \longrightarrow P(Y) = {}^{n_2}C_y \, p^y q^{n-y}$

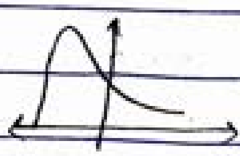
$$\boxed{X + Y \sim B(n_1 + n_2, p)}$$

$$\boxed{P(X+Y) = {}^{n_1+n_2}C_{x+y} \cdot p^{x+y} \cdot q^{n-(x+y)}}$$

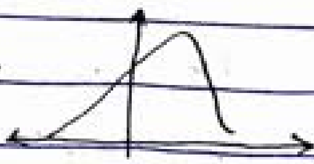
$B(n, p) \rightarrow$ Binomial distribution where n, p are two parameters.

Note:- $P(x \geq 1) = 1 - P(x=0)$.

- $P \ngtr 0.5 \longrightarrow$ +ve skewed.
  skewness towards right



- $P < 0.5 \longrightarrow$ -ve skewness.
  skewed towards to left.



* Poisson Distribution:

Poisson distribution use when -

$n \geq 50$, $p \leq 0.05$

(repeation of succes is two low & Independent on previous succes).

$$\boxed{P(x) = \frac{e^{-m} m^{x}}{x!}}$$ where, $\boxed{m = np}$

n → no. of trial, p → prob. of succes.

$\boxed{e = \text{exponent} = 2.71828.}$

Imp. points.

- $f(0) + f(1) + f(2) + \ldots\ldots = 1.$
- Poisson dist. also known as $\boxed{\text{uniparametric}}$ distribution. as it is characterized by only one parameter. $\boxed{m}$.
- mean of Poisson dist. $\boxed{(\mu) = m}$
- Variance of Poisson dist. $\boxed{(\sigma^2) = m}$

$\Rightarrow \boxed{\mu = \sigma^2 = m}$

- S.D $\boxed{(\sigma) = \sqrt{m}}$
- Mode. (μ_0)

if m is non-integer → greatest integer in m.
if m is an integer → m & m-1

- Poisson distribution always positively skewed ($P<0.5$)

- $\left.\begin{array}{l} X \sim P(m_1) \\ Y \sim P(m_2) \end{array}\right\} \Rightarrow \boxed{X+Y \sim P(m_1+m_2)}$

* Application of Poisson distribution –

- no. of printing mistakes per page of a large book.
  $n \to \infty$ & $p \to 0$.

- no. of road accidents on a busy road $\boxed{\text{per minute}}$
  $p \leq 0.005$ time interval.

- no. of shoe laces $\boxed{\text{per minute}}$ at shoe store.
  $p \to 0$.

- no. of demands $\boxed{\text{per minute}}$ for health center and so on.
  $p \to 0$

Calculator trick. $(x)^{-m}$

1. x
2. '$\sqrt{\ }$' 12 times
3. -1
4. $\times(-m)$
5. $+1$
6. '$x=$' 12 times.

ex → $(2.71828)^{-1.2}$. $= 0.30109$

↓

→ 2.71828.

↓

→ '$\sqrt{\ }$' 12 times

↓

→ -1

↓

→ $\times -1.2$

↓

→ $+1$

→ $x=$ 12 times.

* Normal Distribution / Gaussian distribution.

→ A continuous random variable is defined in term of its probability density function.

$f(x) > 0. \qquad x \in (-\infty, \infty).$

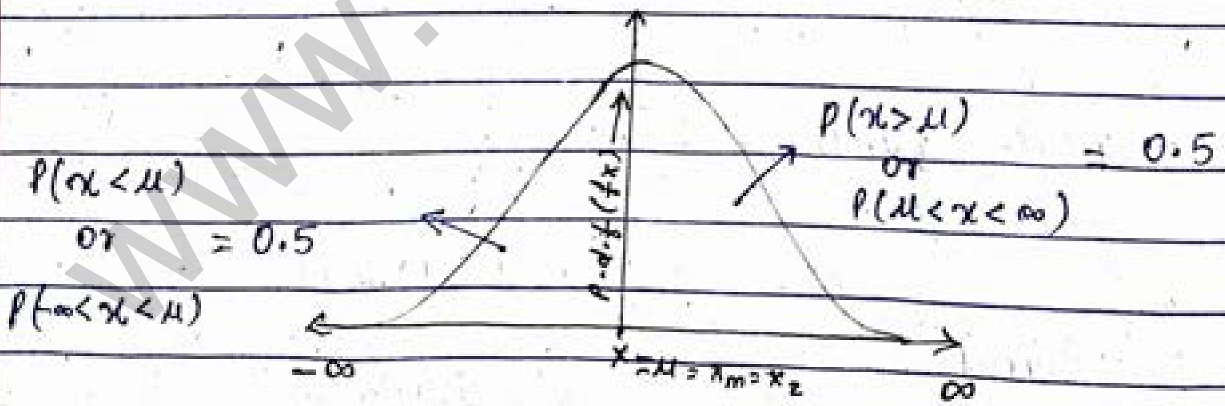
$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

→ A continuous random variable x is defined to follow normal distribution with parameters μ and σ^2.

→ Density function of a normal dist. ⟶ Normal density function.

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \quad \text{or} \quad \boxed{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

Imp. points ⇒



The area b/w $-\infty$ to μ = area b/w μ to ∞ = 0.5.

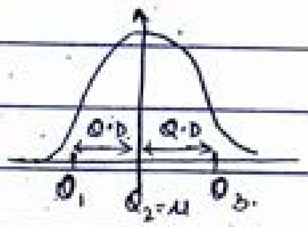
Properties of Normal distribution.

- $\int_{-\infty}^{\infty} f(x) = 1$

- Mean, median, mode of a normal dist will be μ

- Also known as $\boxed{\text{Biparametric distribution}}$ as it is characterized by two parameters μ & σ^2
  $\boxed{X \sim N(\mu, \sigma^2)}$

- Mean deviation $= \boxed{0.8\,\sigma}$ / 0.8 standard deviation.

- 

$Q_1 = Q_2 - Q.D \longrightarrow \boxed{Q_1 = \mu - Q.D}$

$Q_3 = Q_2 + Q.D \longrightarrow \boxed{Q_3 = Q.D + \mu}$

$$\boxed{Q.D = 0.675\, S.D}$$

Standard Normal Distribution.

$\mu = 0, \sigma = 1, \sigma^2 = 1$

$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}} \quad -\infty < z < \infty.$$

Standard normal density function.

- A random variable that has a normal dist. with $\mu = 0$ & $\sigma = 1$ is said to have standard normal probability dist.

Page
Date

Cumulative Distribution function

$$\boxed{\phi(x) = P(X<x) = \int_{-\infty}^{x} f(x)dx}$$

$\phi \rightarrow$ phi.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

$x \Rightarrow z \quad \mu = 0, \sigma = 1$

$$\phi(z) = P(Z<z) = \int_{-\infty}^{z} f(z)\,dz$$

Biometrika table.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty$$

- $P(x < a)$

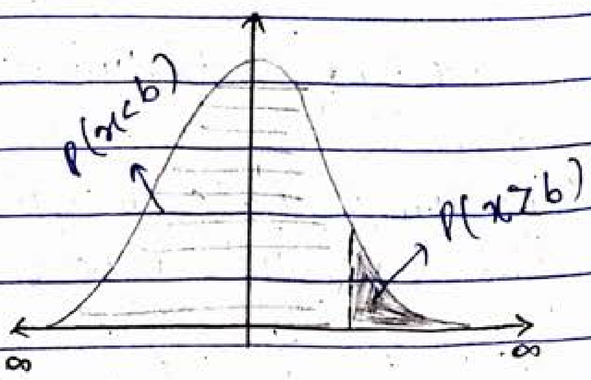
$\downarrow$

$P\left(\frac{x-\mu}{\sigma} < \frac{a-\mu}{\sigma}\right)$ $\left(\text{let, } Z = \frac{x-\mu}{\sigma}, K = \frac{a-\mu}{\sigma}\right)$

$$\boxed{P(Z<K) \rightarrow \phi(K)}$$

Note: $\boxed{P(x \leq a) = P(x < a)}$ as x is continuous.

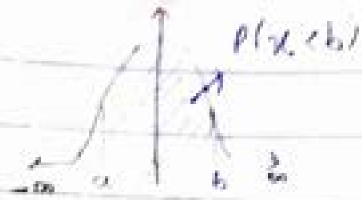
- $\boxed{P(x>b) = 1 - P(x<b)}$



$>$ → right handside area.
$<$ → left " " "

$$P(x<b) + P(x>b) = 1$$

- $\boxed{P(a<x<b) = P(x<b) - P(x<a)}$



let, $z = \frac{x-\mu}{\sigma}$ $\quad = P\left(\frac{x-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) - P\left(\frac{x-\mu}{\sigma} < \frac{a-\mu}{\sigma}\right)$

$K_1 = \frac{b-\mu}{\sigma}$ $\quad = P(z<K_1) - P(z<K_2)$

$K_2 = \frac{a-\mu}{\sigma}$ $\quad = \phi(K_1) - \phi(K_2)$

- $\boxed{P(x<-b) = P(x>b)}$

Formulas:- * $P(z>K) = 1 - P(z<K)$

* $P(z<-K) = P(z>K) = 1 - P(z<K)$

* $P(K_1 < z < K_2) = P(z<K_2) - P(z<K_1)$

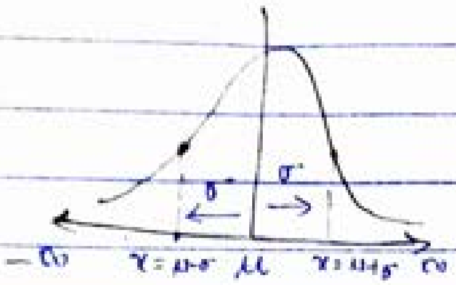
* $P(z>-K) = P(z<K)$.

* $\phi(0) = P(z<0) = P(z>0) = 0.5$

* $\phi(1) = P(z<1) = 0.8413$

* $\phi(2) = P(z<2) = 0.9772$

Properties of Normal distribution

- 

$x = \mu - \sigma$

$x = \mu + \sigma$.

(Two points of inflexion)

- $\text{mean}(\mu) = \mu_1 + \mu_2$.
- $S.D(\sigma) = \sqrt{\sigma_1^2 + \sigma_2^2}$
- $\left.\begin{array}{l} X \sim N(\mu_1, \sigma_1) \\ Y \sim N(\mu_2, \sigma_2) \end{array}\right] \rightarrow X + Y \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Properties of Standard Normal Variate.

- Z has mean, median, mode all $= 0$.
- S.D of Z is 1
- Mean deviation & Quartile deviation is 0.8 and 0.675
- The standard normal distribution is symmetrical about $z = 0$.
- Point of inflexion of prob. curve of standard normal dist are-

$\mu - \sigma = 0 - 1 = -1$

$\mu + \sigma = 0 + 1 = +1$

- The two tails of the standard normal curve never touch the horizontal axis.

- $P(Z > z_p) = p$.

  & $P(Z < z_{1-p}) = p$.

  i.e $P(Z < -z_p) = p$.

  $z_{0.005} = 2.58 \longrightarrow P(Z > 2.58) = 0.005$.

Page:
Date:

Correlation and Regression

Correlation – It establishes the relationship between two or more variables.

* Bivariate data

↓

one of the statistical analysis where two variables are observed

↓

One variable → dependent, other variable → independent.

Bivariate data

↓

Marginal distribution	Conditional distribution
⇓	⇓
X & Y → Bivariate data	X & Y bivariate data.
↓	↓
X ka distribution with its frequency Y ka distribution with its freq.	X ki condition per Y ka distr. Y ki condition per X ka distribution.
⇓	↓
They both are independent of other variable.	Both (X & Y) are dependent on each other variables.

* Positive and Negative Correlation / Uncorrelated.

- Positive correlation → two variable ⇒ positive → If an ↑ in one variable corresponds to ↑ in another variable.

$$\begin{bmatrix} X & Y \\ \uparrow & \uparrow \\ \downarrow & \downarrow \end{bmatrix}$$

- Negative correlation → two variable → negative if an ↑ in one variable corresponds to ↓ in another variable.

$$\begin{pmatrix} X & Y \\ \uparrow & \downarrow \\ \downarrow & \uparrow \end{pmatrix}$$

Page ______

Date ______

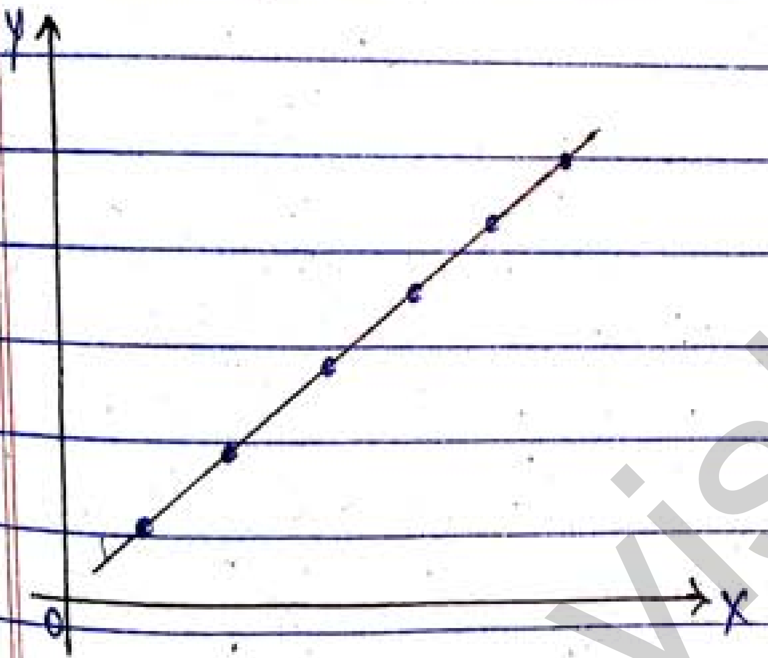
* Methods of Correlation

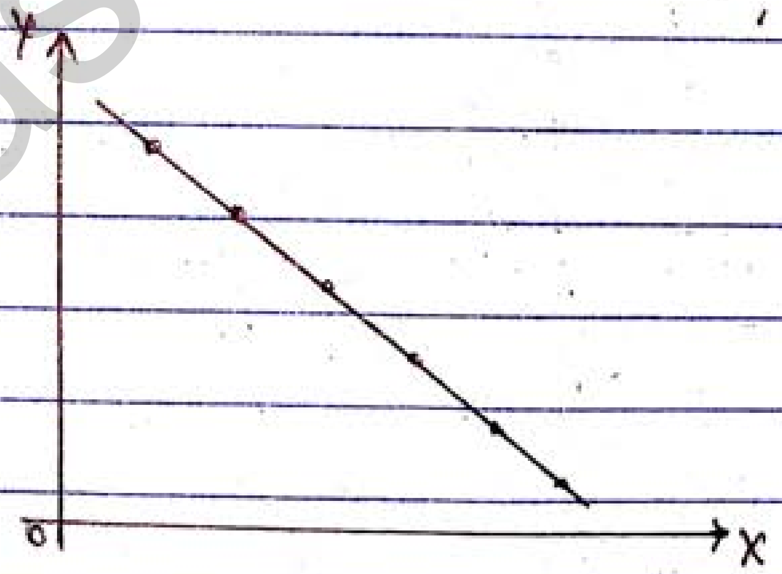
- Scatter Diagram method
- Karl Pearson's Coefficient of Correlation.
- Spearman's Rank correlation method.
- Co-efficient of concurrent deviation.

* Scatter Diagram Method.

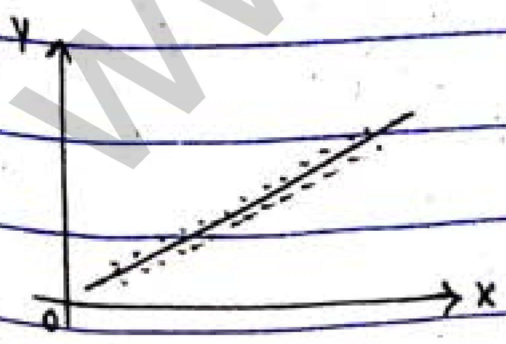
- Perfect +ve correlation.
(Left bottom to Right top)
$r = 1$

- Perfect Negative correlation
(Left top to right bottom)
$r = -1$

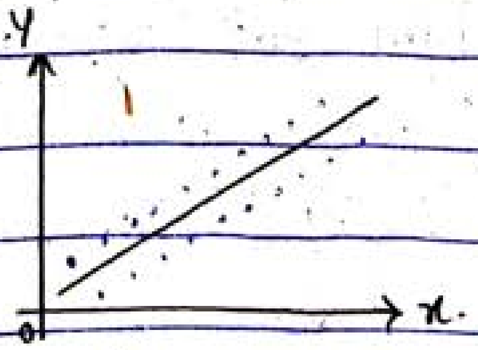




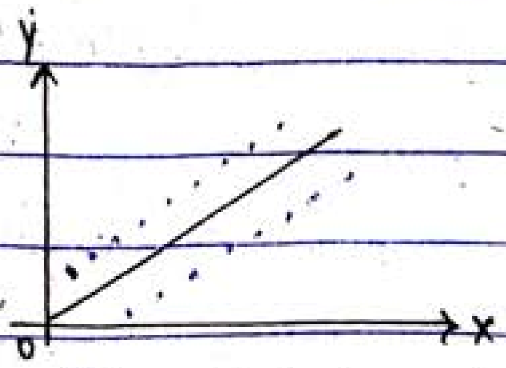
- Positive Correlation.
$(0 < r < 1)$



(Highly +ve correlation).

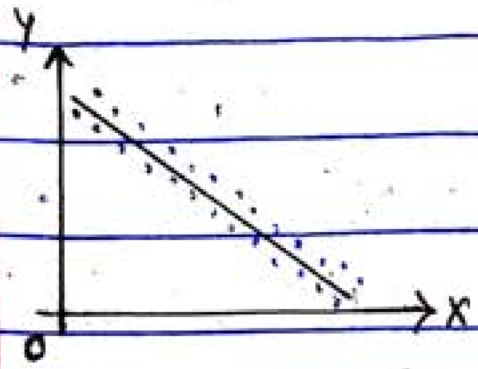


(moderately +ve correlation).

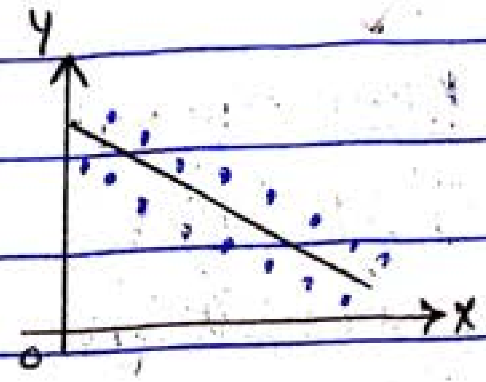


(Less +ve correlation).

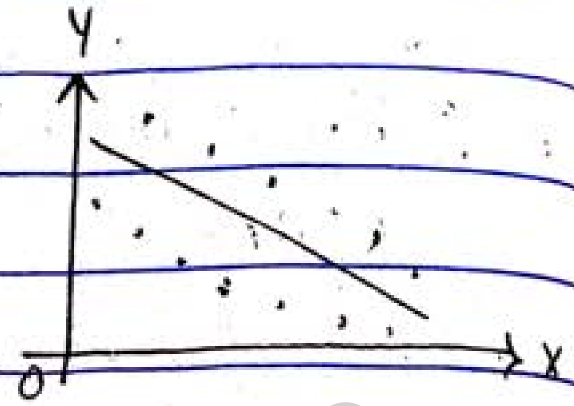
- **Negative Correlation**
  $(-1 < r < 0)$



(Highly -ve correlation)

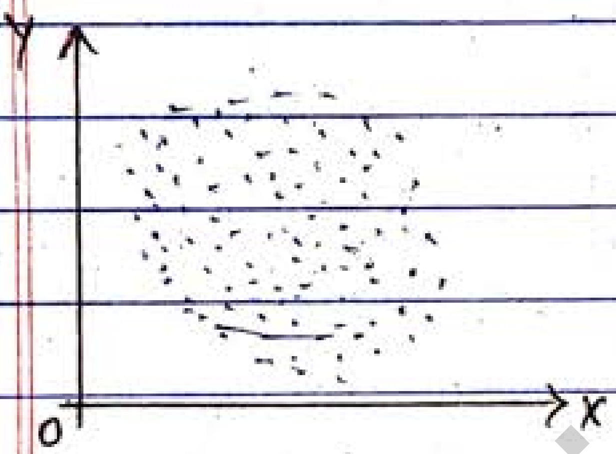


(moderately -ve correlation).

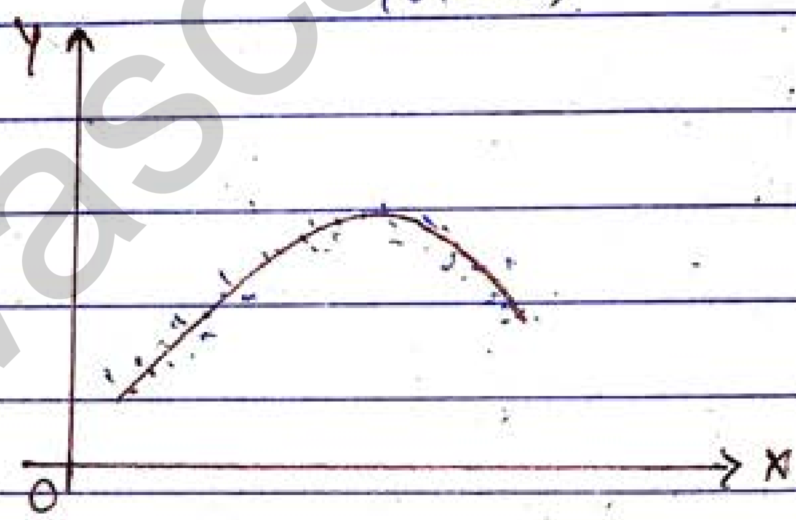


(Less -ve correlation)

- **No Correlation**
  $(r = 0)$



- **Curvilinear correlation**
  $(r = 0)$



The Correlation Coefficient (r).

↓

quantitative measure of strength in the linear relationship b/w two variables.

↓

range $(-1 \leq r \leq 1)$.

↓

if two variables have no linear relationship, the correlation b/w them is 0

↓

the more correlation differs from zero, the stronger the linear relationship b/w two variables.

Date

r

$-ve \rightarrow -ve$ linear correlation

$0 \rightarrow$ No linear correlation.

$+ve \rightarrow +ve$ linear correlation.

Covariance.

↓

average of product of deviation of x from its mean & Y variable from Y's mean.

↓

$$\boxed{Cov(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n}}$$

OR

$$\boxed{Cov(X,Y) = \frac{\Sigma XY}{n} - \bar{X}\cdot\bar{Y}}$$

★ Karl Pearson's Coefficient of Correlation

denoted by 'r'

$$\boxed{r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}} \quad \text{or} \quad \boxed{r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2 \cdot \Sigma(Y-\bar{Y})^2}}}$$

$$\boxed{r = \frac{n\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2)(n\Sigma Y^2 - (\Sigma Y)^2)}}}$$

$$\boxed{r = \frac{N\Sigma f_i u_i v_i - \Sigma f_i u_i \cdot \Sigma f_i v_i}{\sqrt{(N\Sigma f_i u_i^2 - (\Sigma f_i u_i)^2)(N\Sigma f_i v_i^2 - (\Sigma f_i v_i)^2)}}}$$

where, $N = \Sigma f$

$X \Rightarrow U = \frac{X - A_x}{n}$

$Y = V = \frac{Y - A_y}{n}$

Page ____ Date ____

Characteristics of the Correlation coefficient (r).

- r is independent of change of origin and scale.
- if X & Y replaced by variable $U = aX+b$ & $V = cY+b$.

$r(U,V) = \pm 1 \cdot r(X,Y)$

$r(\underbrace{aX+b}_{U}, \underbrace{cY+d}_{V}) = \dfrac{(a \times c)}{|a||c|} r(X,Y)$

a	c	
+	+	$r(U,V) = r(X,Y)$
−	−	
+	−	$r(U,V) = -r(X,Y)$
−	+	

Q. The correlation coefficient b/w X & Y is 0.4; what is the correlation coefficient b/w 2X and (−Y)?

$U = aX+b$, $V = aY+b$.

$\Rightarrow$ $U = 2X$

$V = -Y = -1 \times Y$.

a	c
+2	−1

$r(U,V) = -r(X,Y)$.

$\boxed{r(U,V) = -0.4}$ Ans.

Note-1. Bivariate frequency table having $(p+q)$ classification (p → rows, q → column) the total no. of cells is pq.

2. Spurious Correlation → correlation b/w two variables having no causal relation.

Date

★ **Spearman's Coefficient of Rank Correlation. (ρ)**

↓

denoted by (ρ)

↓

range from $-1 \leq \rho \leq 1$

↓

Value of 1 indicates perfect association for identical ranking

&

value of -1 indicates perfect association for reverse ranking.

formula → $\boxed{\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)}}$ where, D = deviation of ranks.

n = total no. of pairs.

When Ranks are equal.

↓

$$\boxed{\rho = 1 - \frac{6\left(\Sigma D^2 + \frac{\Sigma(t^3 - t)}{12}\right)}{n(n^2-1)}}$$

t = tie count.

★ **Coefficient of Concurrent Deviations. (r_c)**

$$\boxed{r_c = \pm\sqrt{\pm\frac{2c - m}{m}}}$$

where, m = no. of pair of sign (m = n-1)

c = no. of concurrent pairs [same sign wale pair]

Regression

it check how much one variable is dependent on another variable.

Dependent variable

variable whose value is to be predicted

also known as explained variable.

Independent variables.

variables which are used to ~~the~~ predict the values of a dependent variable.

also known as explanatory variables.

⇒ **Simple Regression analysis & Simple linear analysis.**

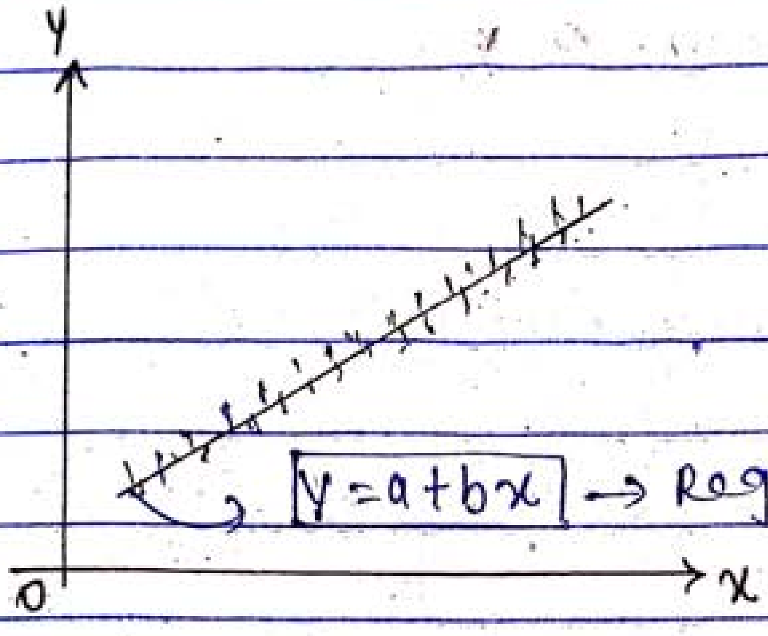
- when the relationship b/w the dependent variable & the independent variable is linear, the technique for prediction is called simple linear analysis.

equation → $\boxed{y = a + bx}$ (Y on x → y is dependent on x).

- Regression analysis of only two variables, a dependent & an independent variable is called simple regression analysis.

Method of Least square.

If a line of best fit approximating the given data has the equation then the method of least square requires that we must determine constants a & b, as to minimize.



$\boxed{Y = a + bx}$ → Regression line → is a line which gives minimum sum of vertical distance from points.

$\Sigma(Y - (a + bx)) = \text{minimum}$.

Y on X regression line

equation → $\boxed{Y = a_{yx} + b_{yx} x}$

x → independent
Y → dependent

a = regression constant
b = regression coefficient.

$$\boxed{\begin{aligned} \Sigma y &= na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}}$$
→ Normal equation of Y on X regression line.

By using normal eqⁿ —

$Y = a + bx \;|\; Y = a_{yx} + b_{yx} x$

where, $\boxed{b_{yx} = \frac{\text{cov}(X, Y)}{\sigma_x^2}}$ OR $\boxed{b_{yx} = \frac{n\Sigma XY - \Sigma X \cdot \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}}$ OR $\boxed{b_{yx} = \frac{r\sigma_Y}{\sigma_X}}$

Note :- On a line $(Y = a + bx)$ always mean lies.

$\boxed{\bar{Y} = a_{yx} + b_{yx}\bar{x}}$

$\Rightarrow \boxed{a_{yx} = \bar{Y} - b_{yx}\bar{x}}$

$\boxed{Y - \bar{Y} = b_{yx}(x - \bar{x})}$

$\boxed{\text{Slope of the line of regression of Y on X} = b_{yx}}$

Page
Date

X on Y (x is dependent on y).

$\boxed{x = c + dy}$

$C = a_{xy} \rightarrow$ regression ~~coefficient~~ constant of x on y.

$d = b_{xy} \rightarrow$ regression ~~constant~~ coefficient of x on y.

$$\boxed{x = a_{xy} + b_{xy}\, y}$$

$$\boxed{b_{xy} = \frac{\text{cov}(x,y)}{\sigma_y^2}} \quad \text{or} \quad \boxed{b_{xy} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum y^2 - (\sum y)^2}} \quad \text{OR} \quad \boxed{b_{xy} = \frac{r\sigma_x}{\sigma_y}}$$

$\bar{x}, \bar{y} \rightarrow \boxed{\bar{x} = a_{xy} + b_{xy}\bar{y}}$

$$\boxed{a_{xy} = \bar{x} - b_{xy}\bar{y}}$$

$$\boxed{\text{Slope of regression line } x \text{ on } y = \frac{1}{b_{xy}}}$$

$$\boxed{x - \bar{x} = b_{xy}(y - \bar{y})}$$

Properties of Regression Coefficient

Property 1 -

r	b_{yx}	b_{xy}.
+ve	+ve	+ve
−ve	−ve	−ve.

Property 2 - r is G.M of b_{yx} & b_{xy}.

$$\boxed{r = \pm\sqrt{b_{yx} \cdot b_{xy}}}$$

$-1 \leq r < 1$

$-1 \leq \sqrt{b_{yx} \cdot b_{xy}} < 1$

$\boxed{b_{yx} \cdot b_{xy} \leq 1}$

(if $b_{xy} > 1 \longrightarrow b_{yx}$ will be less than 1 or vice versa).

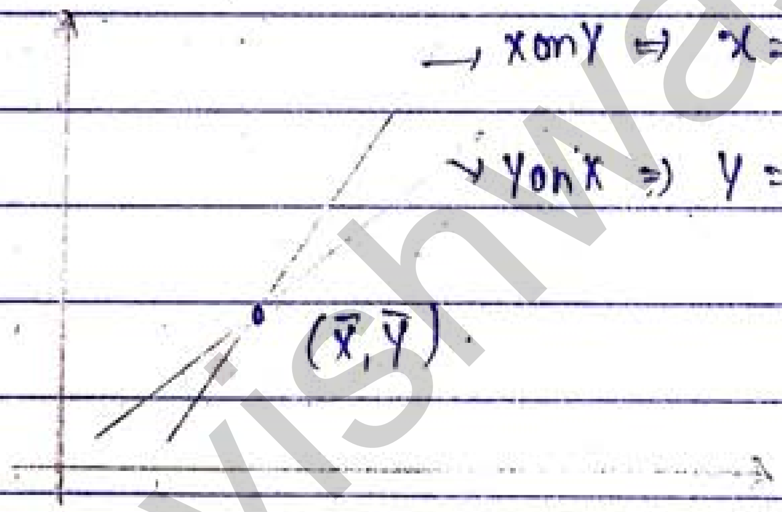
Property 3 → If one of the regression coefficient is greator than unity(1), the other must be less than the unity (1).

$$\text{if } b_{yx} \geq 1 \Rightarrow b_{xy} \leq 1 \quad \text{or} \quad b_{xy} \geq 1 \Rightarrow b_{yx} \leq 1.$$

Property 4 → The two lines of regression intersect at the point $(\bar{x}, \bar{y})$ where x and y are the variables.

$\rightarrow$ x on y $\Rightarrow x = a_{xy} + b_{xy} y$.

$\rightarrow$ y on x $\Rightarrow y = a_{yx} + b_{yx} x$.



Property 5 → The regression coefficients are independent of change of origin but not for scale.

$$U = px + q \qquad V = ry + s$$

$$\boxed{b_{vu} = \left(\frac{r}{p}\right) b_{yx}} \quad \Bigg| \quad \boxed{b_{uv} = \left(\frac{p}{r}\right) b_{xy}}$$

Page
Date

Coefficient of determination

ratio between the explained variance to total variance.

denoted by r^2.

Formula - COD = $\boxed{r^2 = \frac{\text{Explained Variance}}{\text{total variance}}}$.

Coefficient of non-determination = $\boxed{1 - r^2}$ = $\boxed{\frac{\text{unexplained variance}}{\text{total variance}}}$

Imp. Ques: Out of the following two regression lines, find the lines of regression of x on y

$2x + 3y = 7$ and $5x + 4y = 9$

$\frac{x \text{ coefficient}}{y \text{ coefficient}} = \frac{2}{3}$ $\quad$ $\frac{x \text{ coeff}}{y \text{ coeff}} = \frac{5}{4}$

y on x $\leftarrow$ $\frac{2}{3} < \frac{5}{4}$ $\rightarrow$ x on y

(Jo बड़ा है वो x on y hai), & jo छोटा हो वो y on x line h)

x on y $= 5x + 4y = 9$

y on x $= 2x + 3y = 7$.

Page
Date

Index Number

⇓

statistical measures designed to show changes in a variable or a group related variable with respect to time, geographical location, income etc.

A series of numerical figures which show the relative position called Index Number

Simple Index Number → One variable
Composite Index " → More than one variable.

Types of Index Number

Price Index No.
⇓
shows Movements in the price level
ex → Inflation, deflation

Quantity IN
⇓
Show movement of Quantity level.
ex → Industrial Production.

Value IN.
⇓
Show movement in value b/w two periods
Value = Price × Quantity
ex → Growth rate of an Economy.

* **Relatives**

Relatives are derived because absolute numbers measured in some unit and are often of little Importance and meaningless in themselves.

1. Price Relative $= \frac{P_n}{P_0}$ or $\frac{P_n}{P_0} \times 100$.

2) Quantity Relative $= \dfrac{Q_n}{Q_0}$

3) Value Relative $= \dfrac{V_n}{V_0} = \dfrac{P_n \times Q_n}{P_0 \times Q_0} = \dfrac{P_nQ_n}{P_0Q_0}$.

<u>Link Relative</u>

Ratio of <u>successive price</u> or quantities. i.e. $\dfrac{P_1}{P_0}, \dfrac{P_2}{P_1}, \dfrac{P_3}{P_2}, \dfrac{P_n}{P_{n-1}}$

Chain Relative

Ratio is taken in <u>respect</u> to base <u>price</u>. i.e. $\dfrac{P_1}{P_0}, \dfrac{P_2}{P_0}, \dfrac{P_3}{P_0}, \dfrac{P_n}{P_0}$.

Methods for constructing Index Number.

- Simple
  - Aggregative
  - Relative
- Weighted
  - Aggregative
  - Relative.

* <u>Simple Aggregative</u>

current year

Simple aggregative price Index $= \dfrac{\text{Sum of all commodity price in C.Y}}{\text{Sum of all commo. price in base yr}}$

$$P_{0n} = \boxed{\dfrac{\Sigma P_n}{\Sigma P_0} \times 100}$$

Page
Date

Demerits of simple aggregative
- first commodity exerts greater influence than the other two because the price of first commodity is higher than other two
- If units are changed then IN will also change.

* Simple Relative or Simple Average

change the actual price for each variable into percentage of the base period.

$$\text{Simple Relative } (P_{on}) = \frac{\sum \frac{P_n}{P_0} \times 100}{n}$$

eg.

Commodity	2010 (P_0)	2015 (P_1)	2020 (P_1)
milk (per l)	50	60	70
Atta (per kg)	10	12	15
Potato (per kg)	20	30	30

⇒

Commodity	2010 (P_0)	2015 (P_1)	2020 (P_1)
milk (per l)	$\frac{50}{50} \times 100 = 100$	$\frac{60}{50} \times 100 = 120$	$\frac{70}{50} \times 100 = 140$
Atta (per kg)	$\frac{10}{10} \times 100 = 100$	$\frac{12}{10} \times 100 = 120$	$\frac{15}{10} \times 100 = 150$
(Potato (per kg)	100	150	150
Aggregate	300	390	440
Index No.	$\frac{300}{3} \times 100$	$\frac{390}{3} = 130$	$\frac{440}{3} = 146.66$

Page
Date

Simple Relative method

Advantage

- Unit free
- All commodity are given equally importance.

Disadvantage

Give undue weight to a commodity which is used in small quantity because the relatives which have no regard to the absolute quantity will give more than what is due from the quantity used.

or

It gives equal importance to each relative.

* Weighted Aggregate Index

- We weigh the price of each commodity by suitable factor often take as the quantity or value weight sold during the base yr. or given year or an average of some yr.
- Indices are shown as %.

Methods —

1) Laspeyres's Index (weight → Q_0) $= \boxed{\dfrac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} \times 100}$

2) Paasche's Index (weight → Q_1) $= \boxed{\dfrac{\Sigma P_n Q_n}{\Sigma P_0 Q_n} \times 100}$

(3) Marshall-Edgeworth Index (weight → $(Q_0 + Q_1)$) $= \boxed{\dfrac{\Sigma P_1 Q_0 + \Sigma P_1 Q_1}{\Sigma P_0 Q_0 + \Sigma P_0 Q_1} \times 100}$

Date

(4) Fisher's Index $= \sqrt{L \times P}$

$$= \sqrt{\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} \times \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_n}} \times 100$$

(5) Bowley's Index $= \frac{L+P}{2}$ or $\left(\frac{\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} + \frac{\Sigma P_n Q_1}{\Sigma P_0 Q_1}}{2}\right) \times 100$

	P_0 2016	Q_0	P_n 2022	Q_n
Commodity	Price	Quantity	Price	Quantity
A	20	60	25	55
B	30	50	20	60

* **Weighted Average of Relative method**

$\Rightarrow$ $\boxed{\frac{\Sigma \frac{P_n}{P_0} \times (P_0 Q_0)}{\Sigma P_0 Q_0}}$ $\Rightarrow$ $\boxed{\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0}}$ same as Laspeyer's method Index Number

* **Chain Index Number**

$$\boxed{CIN = \frac{\text{Link Relative of Current Year} \times \text{Chain relative of previous year}}{100}}$$

Ex.

Year	Price of commodity	Link Relative	Chain Index	Chain Relative
2010	10	$10/10 \times 100 = 100$	100	$10/10 \times 100 = 100$
2012	12	$\frac{12}{10} \times 100 = 120$	$\frac{120 \times 100}{100} = 120$	$\frac{12}{10} \times 100 = 120$
2014	14	$\frac{14}{12} \times 100 = 116.667$	$\frac{116.667 \times 120}{100} = 140$	$\frac{14}{10} \times 100 = 140$

* **Value Index**

⇓

equal to the sum of values of a given year divided by sum of value of base year.

$$\boxed{\frac{\Sigma V_n}{\Sigma V_0} = \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_0}}$$

Index Number

* **Limitations**

- Chances of error due to sampling
- Not give real picture only give broad trend.
- Creat confusion due to many formulas.

* **Usefulness**

- Very useful in deflating (e.g. nominal wages into real).
- Help in framing suitable policies.
- Reveal trend & tendencies in making important conclusion.

* Deflated Value = $\boxed{\frac{\text{Current value}}{\text{Price index of current year}}}$

or

Deflated value = $\boxed{\text{Current value} \times \frac{\text{Base Price}}{\text{Current Price}}}$

Ex.

Cost of living		Salary	
2015	2019	2015	2018
97.5	115	19500	? 23000

$\Rightarrow \frac{19500}{97.5} \times 115 = \boxed{23000}$

Page
Date

* **Shifting Price Index** → Base year must relative recent.

$$\text{Shifting Price Index} = \frac{\text{Original Price Index}}{\text{Price Index of the year on which it has to be shifted}} \times 100$$

* **Splicing**
⇓
Two index covering different bases may be combined into single series by slicing.

$$\text{New Index Shifting} = \frac{\text{Old Index no.} \times \text{Jisme shift}}{\text{Jisse shift}}$$

Ex-

Year	Old Price I.N (1990=100)	Revised Price Index (1995=100)
1990	100	100×100/114.2 ⇒ 87.56
1991	102.30	102.30×100/114.2 = 89.58
1992	105.30	105.30×100/114.2 = 92.2
1993	107.60	107.30×100/114.2 = 94.22
1994	111.90	111.90×100/114.2 = 97.98
1995	114.20	100
1996	102.50×114.20/100 = 117.05	102.50
1997	106.40×114.2/100 = 121.51	106.40
1998	108.30×114.2/100 = 123.67	108.30

* **Issue involved in Construction of Index Number.**

(a) Selection of data (according to the purpose).
(b) Base Period (should be normal).
(c) Selection of weights (should have reasonable influence on the Index).

Page
Date

* (d) Use of Averages (The $\boxed{G \cdot M}$ is better in averaging relatives but for most of indices A·M is used because of simplicity.

(e) Choice of variables (depend on the purpose)

(f) Selection of formula (select appropriate formula).

* Test of Adequacy.

- Unit free
- Time Reversal Test
- Factor Reversal Test.
- Circular Test

• Unit ~~free~~ Test → method of finding index no. should be unit free

Every formula satisfy this test $\boxed{\text{except Simple aggregate}}$

• Time Reversal Test :- It means if periods are reversed and indices are multiplied result into unity.

$$\boxed{P_{01} \times P_{10} = 1}$$

Time Reversal test satisfy $\boxed{\text{only Fisher \& marshall edgeworth}}$

• Factor Reversal Test → This holds when product of Price Index and the quantity index should be equal to the corresponding value Index.

$$\boxed{P_{01} \times Q_{01} = V_{01}}$$

Only $\boxed{\text{Fisher's}}$ follow this test

Page ______

Date ______

- Circular test -

It is satisfied by weighted aggregative with fixed weighted average and simple geometric mean of price relatives.

Note:- Fisher's method is called Best method / Fisher's Ideal Index No. because it follow every test except circular test.