PART B → STATISTICS and PROBABILITY.

① About data

→ Refers to facts and statistics collected together for reference (or) analysis.

→ Data
- → Stored
- → Visualized
- → Measured
- → Analyzed.

② Categories of data

1) Qualitative data
   - → Nominal
   - → Ordinal

   } No numbers.

2) Quantitative data
   - → Discrete
   - → Continuous.

   } Numbers

Nominal
- → No order
- Eg: → Male/Female

Ordinal
- → Basic order
- Eg: → Good / Avg / Bad

Discrete
- → Also called categorial data
- Eg: → No: of students in class.
- → Holds only finite number of possible value.

Continuous
- → Data can hold infinite number of values.
- Eg:- Weight of person (50, 50.1, 50.02)

Note:

Discrete variable → a = Spam / Not spam ;

Continuous variable → a = weight ;

(dependent & independent variable also there).

③ What is Statistics ?

→ Area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.

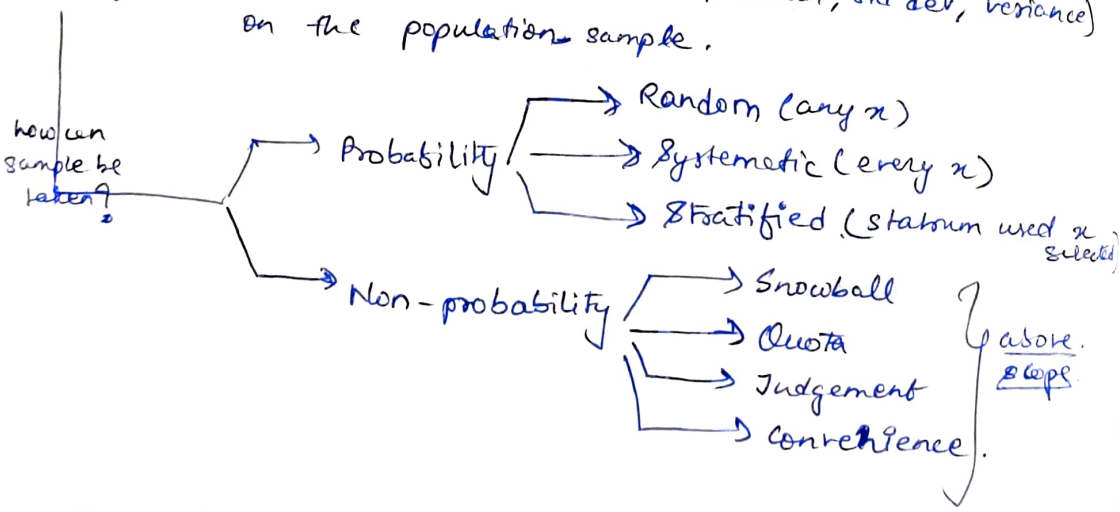
Eg:- i) Data report analysis preparation for a company to identify areas to improve business in.

ii) whether to take a bet or not ? If yes, which side should you be on.

Basic terminologies

→ population , sample , Sampling (taking a sample & inferring).

Sampling → To do infer statistics (mean, median, std dev, variance) on the population sample.



Types of Statistics

Descriptive (Describes data) → Uses data to provide descriptions of the population, either through numerical calculations (or) graphs (or) tables.

Inferential (Infers from data) → Inferences and predictions about a popul -:. Derives conclusions.

(4) Descriptive Statistics

→ Describes (or) summarizes data.

→ Des Stats (categories)
- → Measures of Central Tendency.
- → Measures of Variability / Spread.

Measures of Central Tendency
- Mean (Sum/n)
- Median (order & middle value)
- Mode (most repeated value).

Measure of Spread
- Range
- Inter Quartile Range
- Variance
- Std deviation.

→ (i) **Range** → Max − Min

(ii) **IQR** → Quartiles tell us about spread of dataset by breaking the dataset into quarters.
LIKE median breaks it in half.

→ IQR = $Q_3 - Q_1$ (Each quarter is 25%).

(iii) **Variance** → How much a random var differs from it's expected value. Sq of deviation.

→ $\frac{\sum (x_i - \bar{x})^2}{n}$

[$\bar{x}$ → mean, x_i → element (any), n → total data points].

(iv) Note:

[**Deviation** [Not Std deviation] → Difference between each element from it's mean.

→ $(x_i - \mu)$ (or) $(x_i - \bar{x})$

→ Population and Sample Variance.

(n)	(n−1)
σ^2	s^2

(v) **Std deviation** → Dispersion of a set of data from it's mean.

$\sigma = \sqrt{Var}$

(VI) Information Gain & Entropy

→ Used in random forest and decision Tree, to decide what will be root node.

→ Entropy → Measure of uncertainty present in data.

$$\rightarrow H(S) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

[p_i → event probability
S → set of all instances]

Information Gain → $Gain(A,S) = H(S) - H(A,S)$

[$H(A,S)$ → entropy of attribute A
$H(S)$ → entropy of dataset S]

→ How much 'information' a particular feature/variable gives about final outcome.

→ Entropy $\propto \frac{1}{\text{Significance of variable in decision.}}$

IG ∝ how much info this var will provide.

(VII) Confusion Metrix

→ A table that is often used to describe performance of a classification model, on a set of test data.

∴ It represents a tabular representation of Actual vs predicted values.

→ $\frac{(\text{True +ve}) + (\text{True -ve})}{(\text{True +ve}) + (\text{True -ve}) + (\text{False +ve}) + (\text{False -ve})}$

Predicted \ Actual	Yes	No
Yes	TP	FP
No	FN	TN