30/10/24

Note:- Sampling technique selection always depends on Problem Statement.

Different types of data and Variance:-

Variable:- A variable is a property that can take an any value.

Two kinds of Variable:-

→ Quantative (Numerical) Variables

→ Qualitative (Categorical) variables.

→Quantative Variable:-

A value can be measured and we can perform mathematical operations like (Addition, multiplication, subtraction, Division).

Ex:- mpg, weight, height

gear → Discrete categorical

Numerical → Continous (0-1) (float) ⇒ ∞.
Eg:- Temp

→ Discrete (int) (0-1) ⇒ 1
Eg:- age, year, marks

Discrete → Numerical (Performing calculations)
Eg:- age [35, 27]

→ Categorical (not performing mathematical operations)
Eg:- gear [4, 5, 6], VS, mobile num, aadhar num.

Qualitative Variable :-

A Non-measurable data and Based on characteristics we can derive categorical Variables.

Ex :- gender → male, female, other

Working Type : → IT, NonIT

Blood Group → A+ve, B+ve, AB+ve, O+ve

Variable Measurement Scales :-

We have 4 types of measured variables

- → Nominal Data
- → Ordinal Data
- → Interval Data
- → Ratio

Nominal Data :-

The Categorical Data which are having different classes.

Ordinal Data :-

Order of the Data matters but the values does not matters

Ex :- Marks :- 67, 87, 58, 48, 97

Rank :- ③ ② ④ ⑤ ①

Marks :- 28, 45, 64, 77, 92

Rank :- ⑤ ④ ③ ② ①

Interval Data:-

Order matters and value also matters but natural zero is not present.

Ex:- Sight of Eye (it can't be zero)

Ratio Data:-

The ratio Data can be measured, Order, Equidistinct and have meaningfull zeros. [two zero point]

Ex:- Ratings, no of students, salary, height, age.

31/10/25

Descriptive Statistics:-

→ Measure of Central tendency
→ Measure of dispersion
→ Data Visualization.

Measure of Central tendency:-

→ Mean } Numerical
→ Median }
→ Mode [categorical]

Mean:-

Population Mean:- (μ)

Sample Mean:- ($\bar{x}$)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \frac{\text{Sum of Obs}}{\text{No of obs}}$$

Population mean Sample Mean.

Median:

→ Sort the value either asc (or) desc order.

→ choose the mid value

→ If you get mid (2) values take avg of those 2 values.

Ex: ① [1, 2, 2, 3, 4, 5] , ② [1, 2, 2, 3, 4, 5, 100]

$$\text{mean} = \frac{1+2+2+3+4+5}{6} = \frac{17}{6} = 2.8$$

$$\text{Median} = \frac{2+3}{2} = \frac{5}{2} = 2.5$$

② $$\text{mean} = \frac{1+2+2+3+4+5+100}{7} = \frac{117}{7} = 16.7$$ → outlier (so, far from previous)

median = 3

→ Mean will be affected by outliers where as median won't affect by outliers.

Outliers :- Far Element

→ For null value Imputation - replacement we are using mean and median. [one of the use case]

Mode:- The most repeated values.

Ex: [1, 2, 2, 3, 4, 5, 100]

2.

Example:	Person 1	Person 2
Mon	7:30 AM	8AM
Tues	7:45 AM	11 AM
Wed	8AM	9 AM