Note:
Sampling selection always depends on problem statement.

October - 27:

Types of Variables:

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

Types:
There are two kinds of variables.

1. Quantitative / Numerical variables - A value can be measured and we can perform any kind of mathematical operations like Addition, Subtraction, multiplication, division etc.....

Ex: price, mpg, profit, height, weight, speed.....

Quantitative < Discrete variable (pure integer)
Continuous variable (decimal numbers) (float)

* Discrete variable is also called as discrete categorical data, on this data we won't perform any mathematical operations.

Ex: phone number, pincode, Aadhar number.....

* Discrete Numerical data is the data which is represented in integer form & we do some operations on it.

Ex: Number of children, Age, Number of products, cricket score.

2. Qualitative variables/categorical data - Non measurable data & Based on some characteristics we can derive categorical variables. (String)

Ex: Gender < Male / Female ; weather < Sunny / Rainy / windy

# Variable Measurement Scales:

There are 4 types of measured variables.

1. **Nominal Data:**
The categorical data which are having different classes (unique values) are nominal data.

Ex: Education level, Religion, Emotions, work status.

2. **Ordinal Data:**
Orders of the data matters but values doesn't (Rank).

Ex: Marks

| | | | |
|---|---|---|---|
| 83 | II | 91 | II |
| 75 | III | 82 | IV |
| 86 | I | 85 | III |
| 53 | IV | 94 | I |

3. **Interval Data:**
Order matters & value also matters but natural zero is not present.

Ex: Temperature, Time, Exam with negative marking, Credit score (SAT, JEE etc)

4. **Ratio Data:**
The Ratio data can be measured, ordered, Equidistant & have meaningful zero (true zero).

Ex: Age, Marks, Weight, Speed, salary, Bank balance.

October - 27

Assignment:

Q1. Identify the data types. Specify whether it is continuous / categorical

Activity:-

Number of beating from wife - Ratio — Discrete (Quantitative)

Results of rolling a dice - Nominal — Categorical

Weights of a person - Ratio — Continuous

Weight of Gold - Ratio — Continuous

Distance between two places - Ratio — Continuous

Length of a leaf - Ratio — Continuous

Dog's weight - Ratio — continuous

Blue Color - Nominal — categorical

Number of kids - Ratio — Discrete (Quantitative)

Number of tickets in Indian railways - Ratio — Discrete (Quantitative)

Number of times married - Ratio — Discrete (Quantitative)

Gender (Male or female) - Nominal — Categorical

Q. Classify as Nominal, Ordinal, Interval, Ratio & whether it is continuous / categorical.

Data:

Gender - Nominal, categorical

High school class Ranking - Ordinal, Categorical

Celsius Temperature - Interval, continuous

Weight - Ratio, continuous

Hair color - Nominal, categorical

Socioeconomic status - Ordinal, categorical

Fare Fahrenheit Temperature - Interval, continuous

Height - Ratio, continuous

Type of living accomodation - Nominal, categorical

Level of Agreement - Ordinal, categorical

IQ - Interval, continuous

Sales Figures - Ratio, continuous

Blood Group - Nominal, categorical

Time of Day - Interval, continuous

Time on a clock with Hands - Interval, continuous.

# Descriptive Statistics:

1. Measure of Central Tendency:

- Mean, Median, Mode.

**Mean:** It is further two types.

Population Mean($\mu$)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

~~$x = \{1, 2, 3, 3, 4, 5, 5, 2\}$~~

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\mu = \frac{32}{10}$$

$$\boxed{\mu = 3.2}$$

Sample Mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$x = \{2, 3, 5\}$

$\bar{x} = \frac{10}{3}$

$\boxed{\bar{x} = 3.33...}$

**Median:**

1. Sort the data
2. Pick middle value.
3. If you get mid 2 values, take avg of those 2 values.

$\{1, 2, 3, 4, 5\}$

$\mu = \frac{17}{6}$

$\mu = 2.83...$

median = 3

$\{1, 2, 2, 3, 4, 5, 100\}$

$\mu = \frac{117}{7}$

$\mu = 16.71$

Median : 3.

**Note:** Mean will affect by the outliers where as median won't affect by "outliers".

**Use case of Mean & Median:** Filling the numerical null values we use mean or median.

Null imputation Methods:

1. Ask the client
2. Drop the null values.
3. Fill the null values.

Mode: Most repeated value. Used for categorical data.

For categorical null value place we use mode.

Ex: {Red, Blue, Red, Green, Red, Blue, Green, Red, Green}.