Statistical Measures

Measure of Central Tendency

A measure of central tendency is a statistical measure that represents a typical or central value for a dataset. It provides a summary of the data by identifying a single value that is most representative of the dataset as a whole.

- mean
- median
- mode.

— Mean - mean is the sum of all values in dataset divided by the number of values.

Data = [1, 2, 3, 4, 5]

$x_1\ x_2\ x_3\ x_4\ x_5$

$= \frac{1+2+3+4+5}{5}$

$= \frac{15}{5} = 3$

Sample Data (n)

Sample mean ($\bar{x}$)

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

Population Data (N)

Population mean (μ)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1+x_2+x_3+x_4+x_5}{5}$$

— **Median** :- The median is a middle value in the dataset when the data is arranged in order.

data = [6, 4, 3, 1, 5, 2]

↳ sort = [1, 2, 3, 4, 5, 6]

$$\text{median} = \frac{3+4}{2} = 3.5$$

= [1, 2, 3, 4, 5]

median

— median is always Robust to outlier.

ex - classroom Age in BCA class Final year.

x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}

Age = [19, 21, 20, 19, 21, 22, 20, 21, 19, 20]

$$\text{mean} = \frac{19+21+20+19+21+22+20+21+19+20}{10}$$

$$= \frac{202}{10}$$

mean = 20.2

median = 19, 19, 19, 20, 20, 20, 21, 21, 21, 22

$$= \frac{20+20}{2} = 20$$

Median = 20

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{10}$

Age = [19, 21, 20, 19, 21, 22, 20, 21, 19, 2020]

outlier

$$\text{mean} = \frac{182 + 2020}{10} = \frac{2202}{10}$$

mean = 220.2

median = 19, 19, 19, 20, 20, 21, 21, 21, 22, 2020

min — max

$$= \frac{20 + 21}{2} = 20.5$$

— **Mode** - The mode is the value that appears most frequently in dataset.

→ mode generally use for Categorical Data

Data - Gender = [Male, Female, male, ..., Female] — 100 student

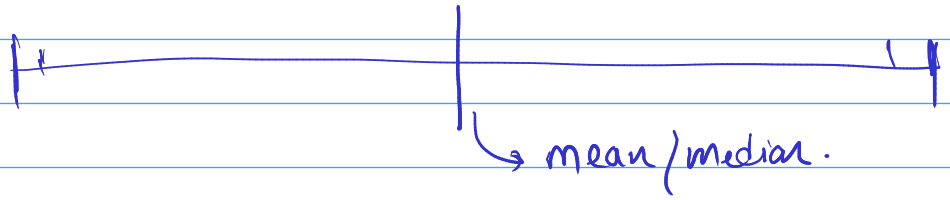
Null, missing

Male = 60

= Female = 60

mode

Measure of Dispersion



A measure of dispersion is a statistical measure that describes the spread or variability of a dataset. It provides information about how the data is distributed around the central tendency (mean, median or mode) of the dataset.

Variance

Variance: The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

Population (σ^2) → sigma

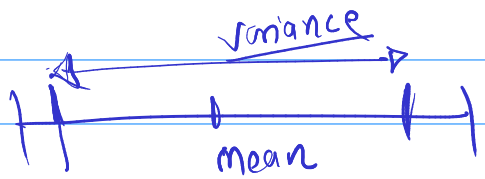
$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample (s^2)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

km = km^2

σ^2 = km^4

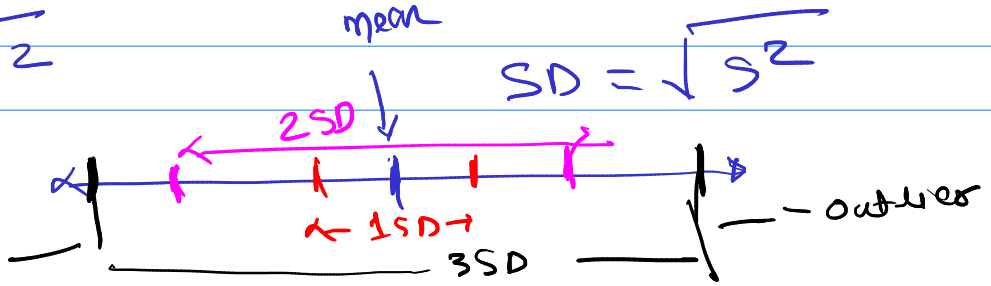


Standard-Deviation

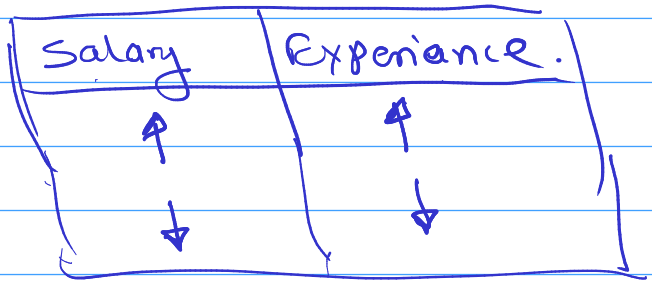
Standard Deviation: The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

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

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



Coefficient of Variation CV



Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

$$CV = \left(\frac{\text{Standard Deviation}}{\text{mean}}\right) \times 100$$

Graph

- Categorical
- Numerical.
  1. Univariate Analysis – (one feature)
  2. Bivariate Analysis (2 feature
  3. Multivariate Analysis (All)

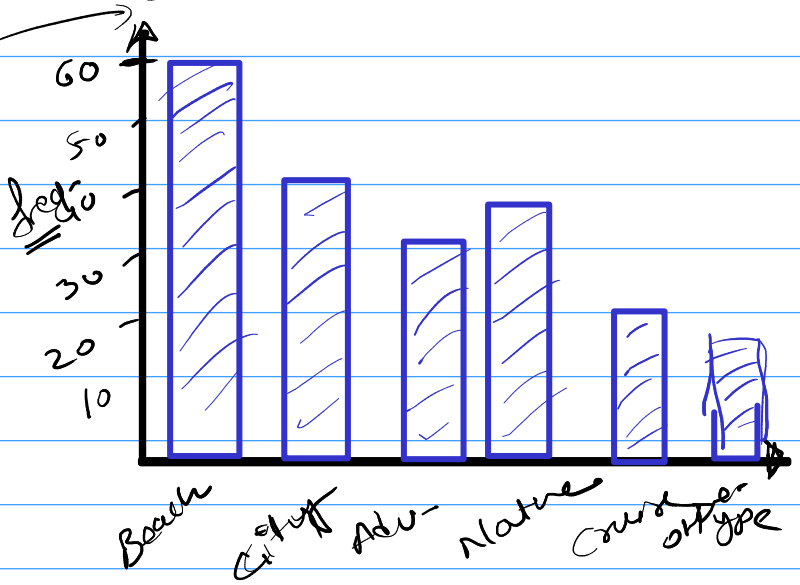
Categorical Data

Frequency Distribution Table.

Ex = 200 people Vacation

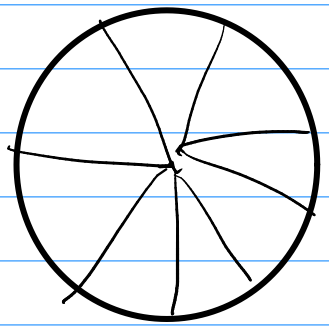
Type of Vac.	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
others	15
	200

BAR chart



Type of Vac.	Frequency	Relative Freq	
Beach	60	0.3	60/200 30%
City	40	0.2	40/200 20%
Adventure	30	0.15	15%
Nature	35	0.175	17.5%
Cruise	20	0.1	10%
others	15	0.075	7.5%
			100%

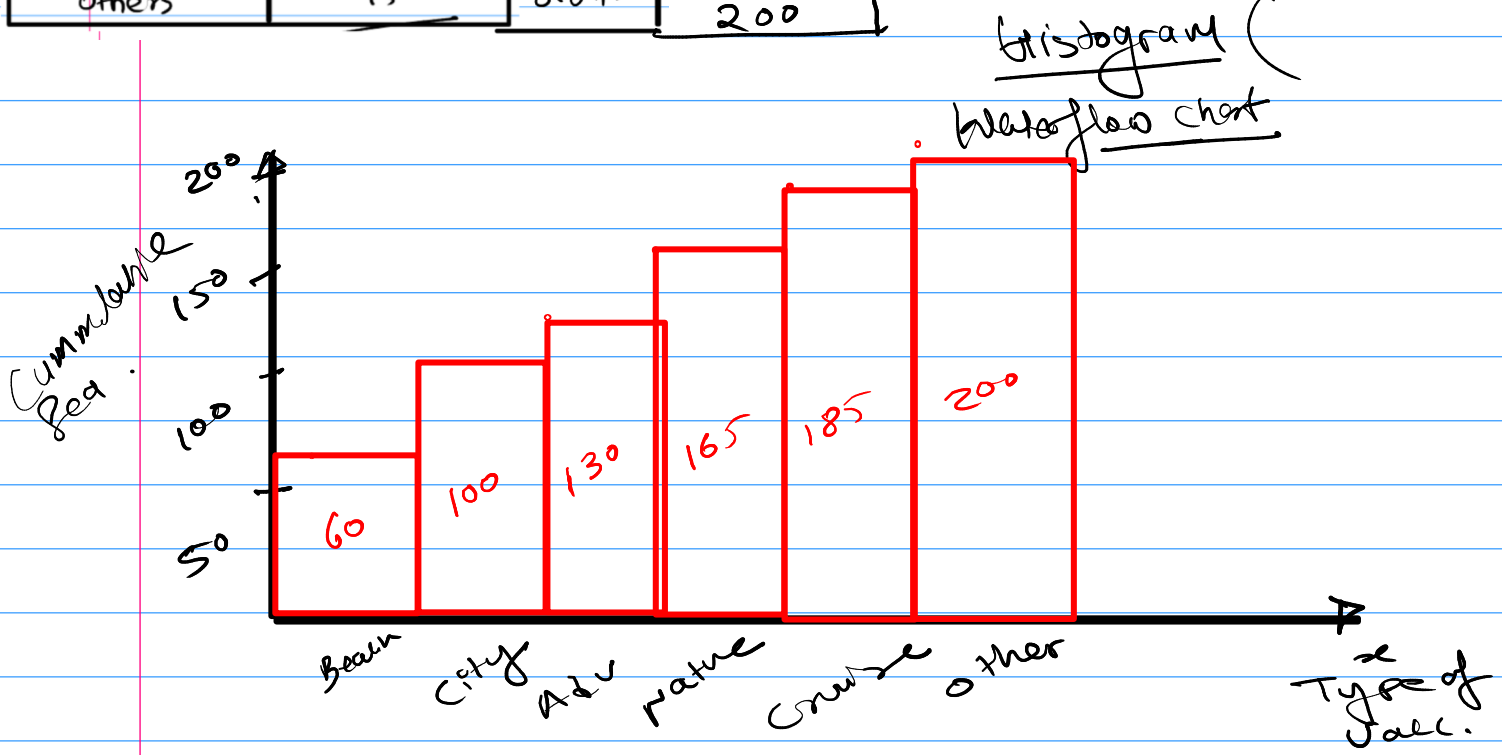
Pie

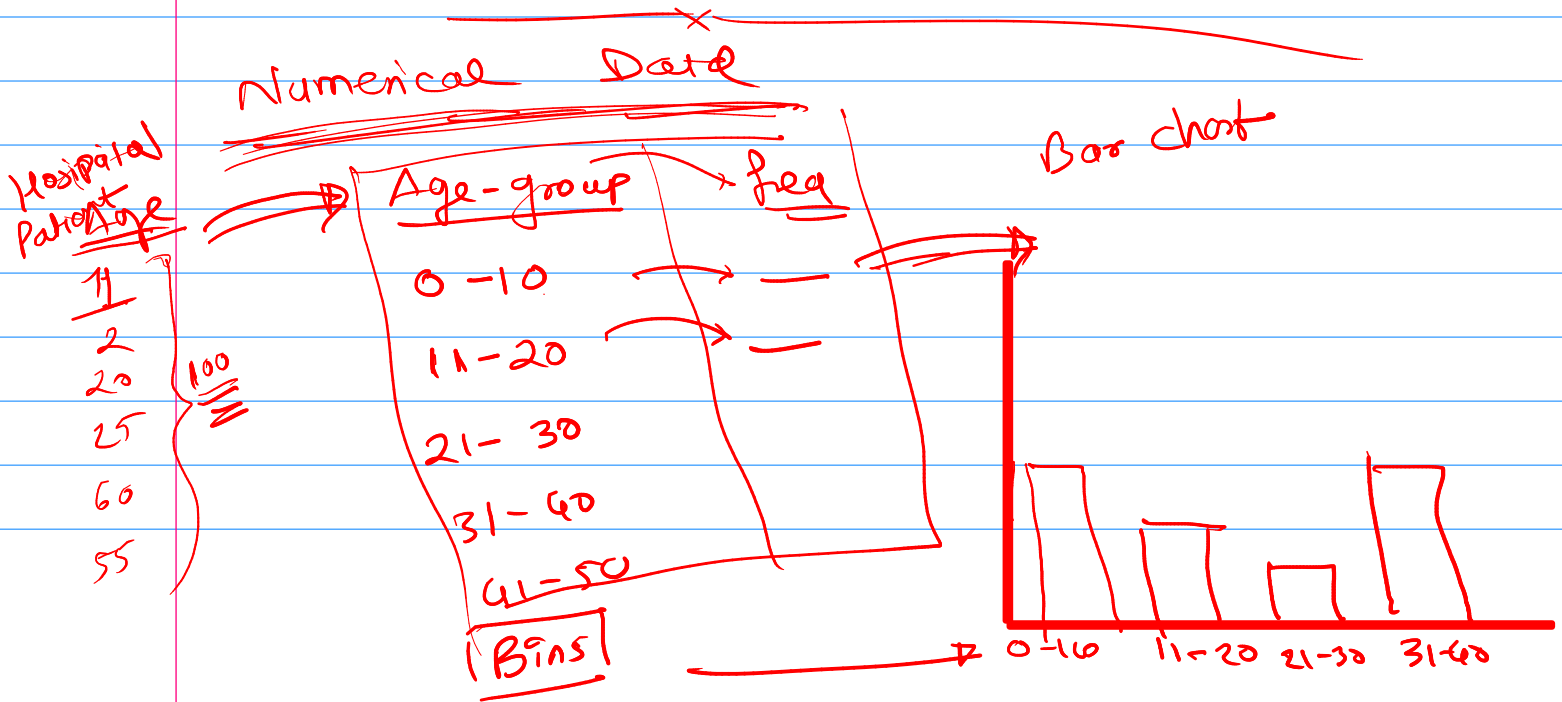


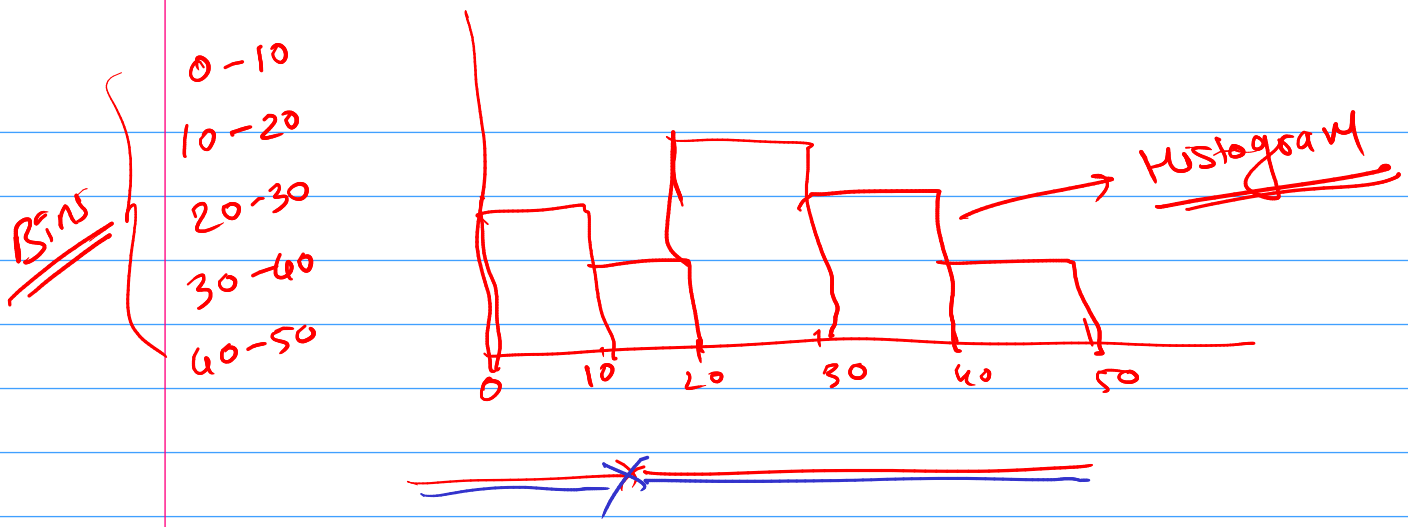
Relative frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

Cumulative frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

Type of Vacc.	Frequency	Relative freq	Cumulative
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
others	15	0.075	200







Category + Category

Titanic Data = Survived Pclass

Death — 0 1

Survived — 1 2

3

Class

Cross-Table

Survived	1	2	3
0	42	31	63
1	71	181	13

Numerical vs Numerical

Scatter plot