Algebric Variable = $x + 5 = 10$

$x = 5$

general Twist

Random Variable = x

Throwing a dice 

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

List Variable = $x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$

List

Statistical Measure's

Measure of Central Tendency

- Measure the central value of Dataset.

Mean — Numerical

The mean is sum of all observation divided by total no. of observation.

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

$x_1\ x_2\ x_3\ x_4\ x_5$

$$\text{mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

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

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Median → Numerical

- The median is the middle value in the dataset, when the data is arranged in order.

$$\text{odd} = \left[\frac{n+1}{2}\right]$$

$$\text{even} = \frac{\left[\frac{n}{2}\right] + \left[\frac{n}{2}+1\right]}{2}$$

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

arrange the data

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

x_1 x_2 x_3 x_4 x_5

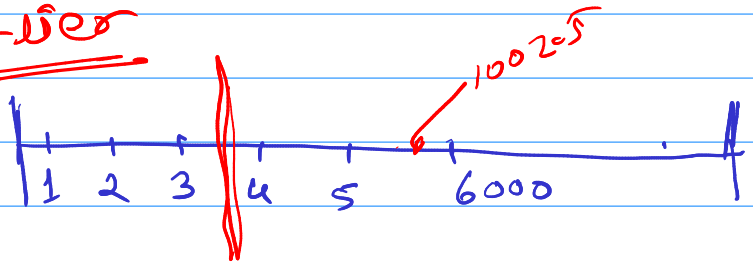
Why median?

— Median is Robust to outlier

eg: data = [1, 2, 3, 4, 5, 6000]

$\bar{x} = \frac{1+2+3+4+5+6000}{6}$ → outlier

mean $= \frac{6015}{6} =$ $\bar{x} = 1002.5$



median = [1, 2, 3, 4, 5, 6] — even

$$\text{median} = \frac{3+4}{2} = \frac{7}{2} = 3.5$$

median = 3.5

Age = [20, 21, 22, 23, 21, 25, 21, 26, 20, 21, ---

(60 student) ---------- 21, 25, 2525, 2020]

2, 0, 2, 0]

first

Last

Mode = The most frequent value is Dataset.

Categorical variable.

Gender = Male, F, M, F, F, M, ==, M, F.

F = 4

m = 4

= There are two mode

mode = M / F

Null — Blank

F = 6

m = 4

mode = F

Weighted Mean - The sum of all the product with its weight, divided by Sum of all weight.

	name	Accuracy	output → House price prediction
AI = Algorithm =	LR :	0.2	10L
	RF :	0.3	15L
	Xgboost :	0.1	8L
		weight	

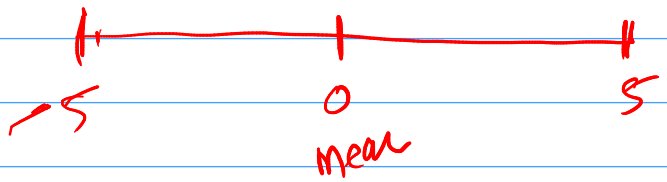
$$= \frac{10L \times 0.2 + 15L \times 0.3 + 8L \times 0.1}{0.2 + 0.3 + 0.1}$$

= Weighted Mean

output	expenses	Weight
	—	Bus
	—	Car
	—	Bike
	—	Train

Measure of Dispersion → Spread. प्रसरण

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.



Range

Range: The range is the difference between the maximum and minimum values in the dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

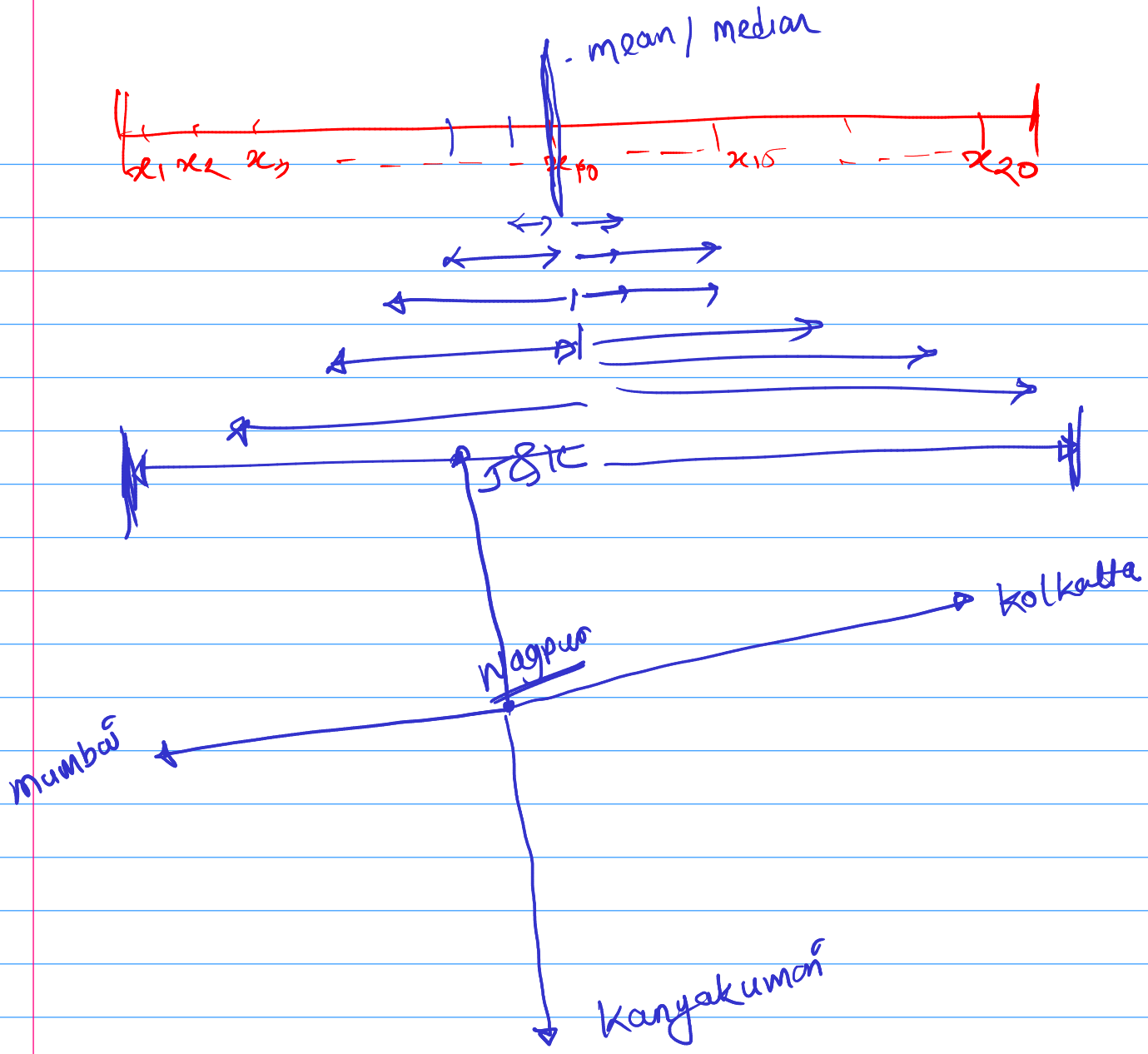
Variance

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.

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

σ → sigma.

why (n-1) ?



Absolute Variance $= |\sigma^2|$ → mod

km → km^2

Laleh → $Laleh^2$

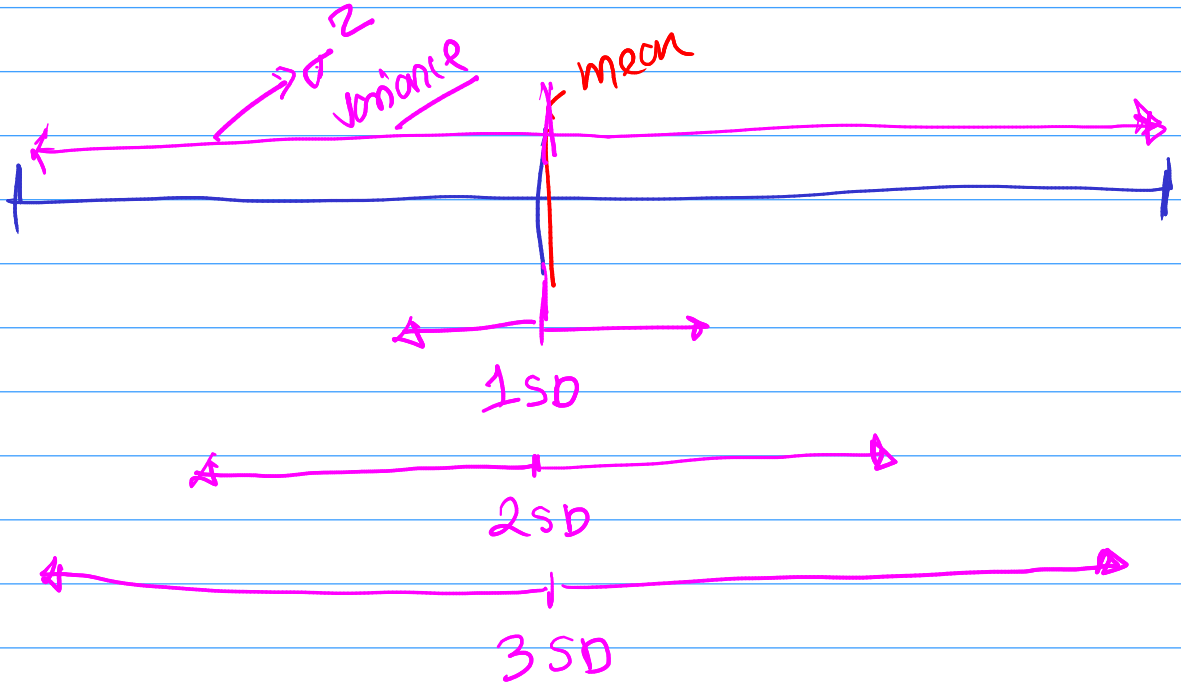
sq.feet → $sq\ feet^2$

3. Standard deviation

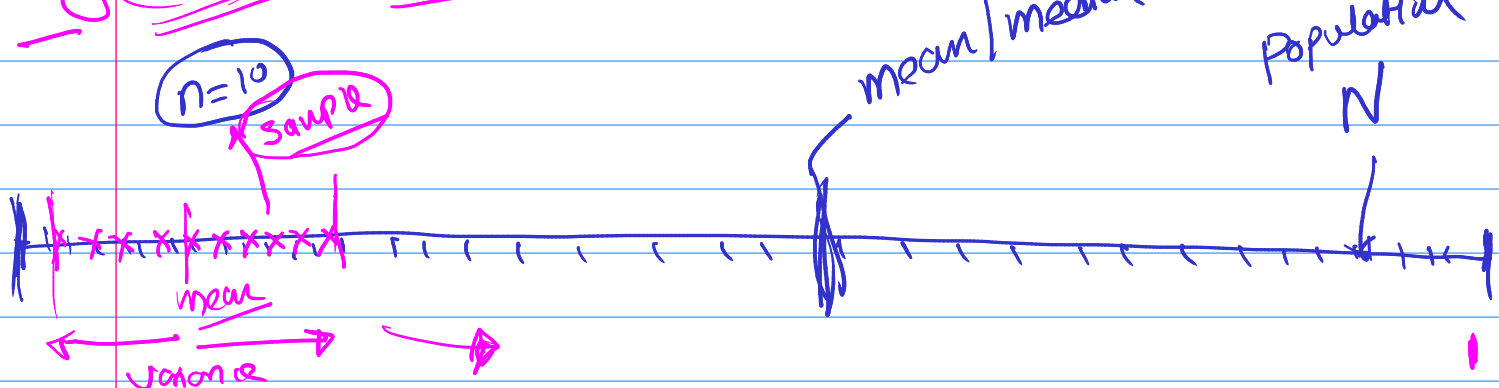
Standard Deviation: The standard deviation is the square root of the variance. It is a widely used measure of dispersion that is useful in describing the shape of a distribution.

(SI unit is same)

$$Sd = \sqrt{s^2} \qquad \sigma = \sqrt{\sigma^2}$$



why (n-1) in sample variance



$\sigma^2 \simeq s^2$ when we devide by (n-1)