Statistics

- Descriptive
  - Central Tendency.
  - Spread of data.
  - Symmetry of data.
- Inferential.
  - population sample
  - Hypothesis testing.

population → 140 + crores

sample → sample data → infer the population

Classroom

Height (in cm)

160
156
160
170
172
165
160

→ Average = $\frac{\text{sum all the observations}}{\text{no. of obs.}}$ = ? =

→ Frequency = 160

Measures of Central Tendency.

① Mean. (i.e Average)

population → μ

sample → $\bar{x}$

Mean = $\frac{\sum_{i=1}^{N} x_i}{n}$ → sum of observations; n → no. of observations.

eg. Salary of people. (in lakh)

10
14
17
18
12
13
$\boxed{101}$

extreme value (outliers)

$\mu = \frac{10+14+17+18+12+13+101}{7}$

$= \boxed{26.4}$

$\mu = \frac{10+14+17+18+12+13}{6}$

$= \cancel{12}\ \boxed{14}$

Concl.

① Mean is easily affected by the presence of extreme points. (outliers)

It is not going to give true indication of data.

② Data treatment (to remove the outliers)

Median

- Middle number

Prog

Salaries Sorted Salaries

↳ Middle number

Prop
1. Data should be sorted.
2. Not affected by outliers

Salaries	Sorted Salaries
10	10
14	12
17	13
18	14
12	17
13	18
101	101

$\frac{N+1}{2}$ ↓ number at the position

$\frac{7+1}{2} = 4$

1. odd observations

~~10~~ ~~12~~ ~~13~~ 14 ~~17~~ ~~18~~ ~~101~~

14 ✓

2. even items

~~9~~ ~~10~~ ~~12~~ $\boxed{13 \quad 14}$ ~~17~~ ~~18~~ ~~101~~

$\frac{13+14}{2} = \frac{27}{2} = 13.5$

$\frac{8+1}{2} = 4.5$ (4^{th} & 5^{th})

Mode → Most frequently occuring value.

0 ~~mod~~ mode
1 "
2 modes (bi-model)
2+ modes (multi mode)

Which one to use?

1.

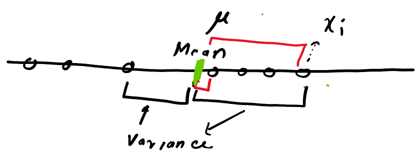
Variability → Measures of Dispersion

population → 100% data is available
sample → S1, S2, S3 .....

μ → Mean

$\bar{x}$ → Sample Mean

Variance → Measure the dispersion of data points around mean.



Distance = $x_i - \mu$

simple datapoint ↓ Mean ↓

1. Population Variance

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

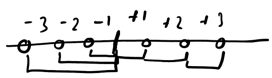
Sum of differences b/w observed value to the population mean divided by total observations

② Sample Variance

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

Summary

① Small is the distance, closer to mean → Data is concentrated around Mean
Higher " " ", away from the mean → More spread



$\sum_{i=1}^{N} \boxed{x_i - \mu} = -3 + (-2) + (-1) + 1 + 2 + 3 = 0$

cf.

1
2
3
4
5

Mean = 3
population variance = 2
sample variance = 2.5

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

$$PV = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

$$= \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2.0$$

$$SV = \frac{10}{5-1} = 2.5$$

population data

1
1
1
2
3
4
5
5
5
5
6
7

$\text{Mean} = \frac{45}{12} = 3.70$

PV = 4.02.

SD

$\boxed{1 \quad 1 \quad 1 \quad 2}$ → Mean = 1.25
Var. = $\boxed{0.18}$ $\boxed{0.25}$ p.v
(1 - 1.25) (n-1)

① STANDARD DEVIATION

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

$\sigma = \sqrt{\text{Variance}}$ = Standard Deviation

Formula

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

population SD

$s = \sqrt{s^2}$

sample SD

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

population SD

sample SD

(2) coefficient of variation (CV)

↳ standard deviation relative to mean.

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

↓

Relative standard deviation

<u>population</u>

$$C_V = \frac{\sigma}{\mu}$$

<u>Sample</u>

$$\hat{C}_v = \frac{s}{\bar{x}}$$