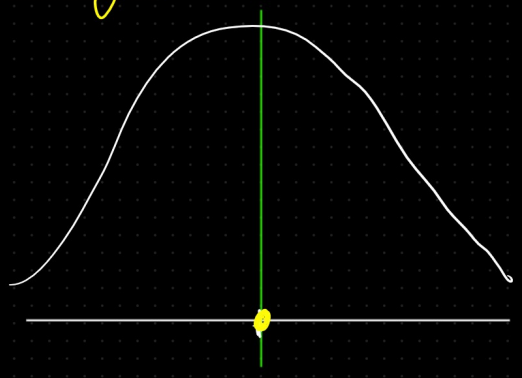
Measure of Central Tendency

1. Mean (Average)
2. Median
3. Mode



Population (N)

Sample (n)

Mean: By adding up all the components and dividing the number of components

$$= \frac{\sum_{i=1}^{N} x_i}{N} \rightarrow \text{population Mean}$$

$$= \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

Population Mean (μ)

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

Sample Mean ($\bar{x}$)

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

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

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2 = \text{Mean}$$

(2) Median : The median is the middle value of a dataset.

$x = \{4, 6, 5, 9, 10, 2, 1\}$

Steps : (1) Sort the Data

$\{1, 2, 4, 5, 6, 9, 10\}$

(2) No of elements → (7)

count = 7

(3) If count is odd

$$M = \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

$$\therefore M = \left(\frac{7+1}{2}\right)^{th} \Rightarrow 4^{th}$$

$$\therefore \boxed{\text{Median} = 5}$$

$$\therefore Y = \{1, 2, (4), (6), 10, 12\}$$

$$\therefore \text{count} = 6$$

$$\text{count} = = \text{even}$$

$$M = \frac{\left(\frac{N}{2}\right)^{th} + \left(\frac{N}{2}+1\right)^{th}}{2}$$

$$= \frac{3^{th} + 4^{th}}{2} = \frac{4+6}{2} = \frac{10}{2}$$

$$\boxed{\text{Median} = 5}$$

Outlier

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

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

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

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

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

$$\bar{x} = \frac{1+2+3+4+5+100}{6}$$

$$= \frac{115}{6} \simeq 19.$$

3 → 19 (outliers)

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

Note: Median is used to find the central tendency when outliers are present.

(3) Mode : The mode is the value that occurs most often.

[Frequency Maximum]

{ 10, 20, 30, 10, 10, 40 }

Mode = 10

NAN → Missing Values → handle

EDA

Age

20

10

30

40

—

50

—

Mean / Median

Mean

Median

Mode

Class

A

B

C

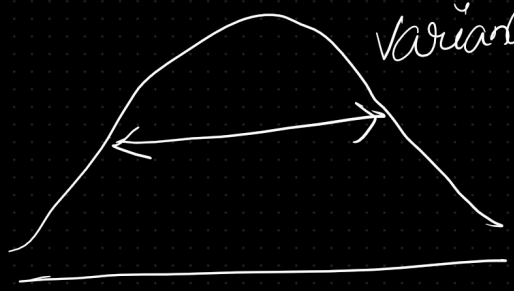
—

—

Mode

Measure of Dispersion [Spread of the data]

1) Variance
2) Standard deviation



Population Variance

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

$x_i \rightarrow$ data points

$\mu \rightarrow$ population mean

$N \rightarrow$ population size

Sample Variance

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

$x_i \rightarrow$ data points

$\bar{x} \rightarrow$ sample mean

$n \rightarrow$ Sample size

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

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
$\bar{x} = 3$		10

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

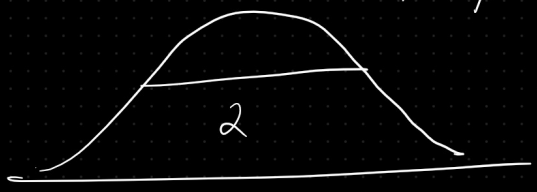
$$= \frac{10}{5-1} = \frac{10}{4}$$

$$S^2 = 2.5$$

$$\sigma^2 = \frac{10}{5} = (2)$$

(2) (2.5)

Dispersion / Spread / Variance



Standard deviation

population std

$$\sigma = \sqrt{\text{variance}}$$

Sample std

$$\text{std} = \sqrt{s^2}$$

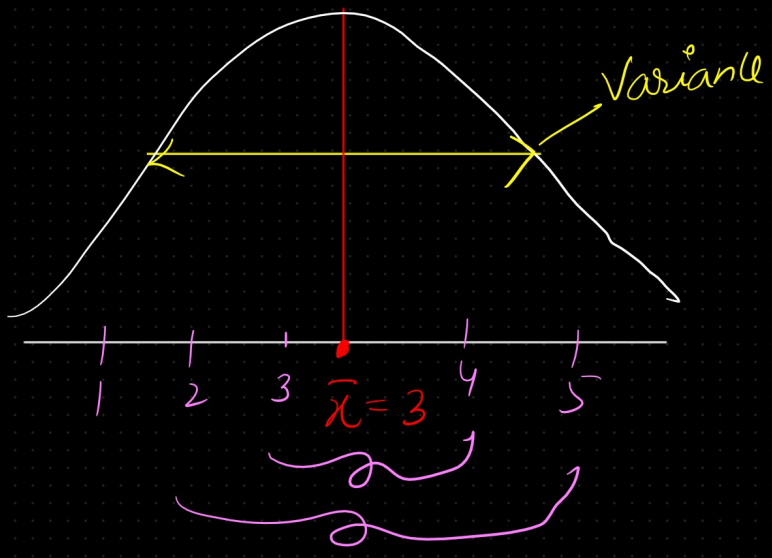
s^2 → Sample Variance

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

$$\bar{x} = 3$$

→ How much individual data points differ from the mean

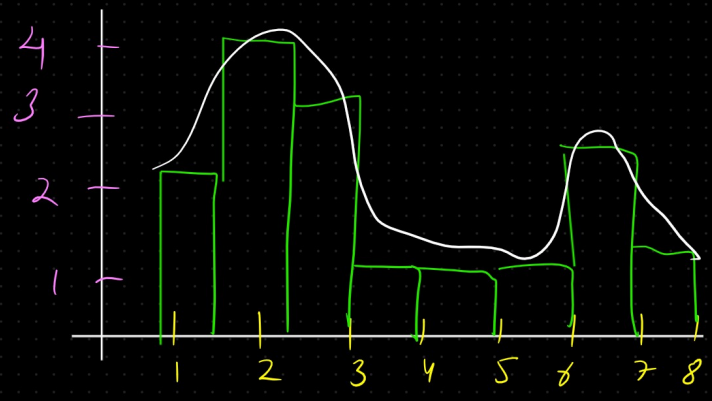
→ It provides a way to express how spread out of the values in a dataset

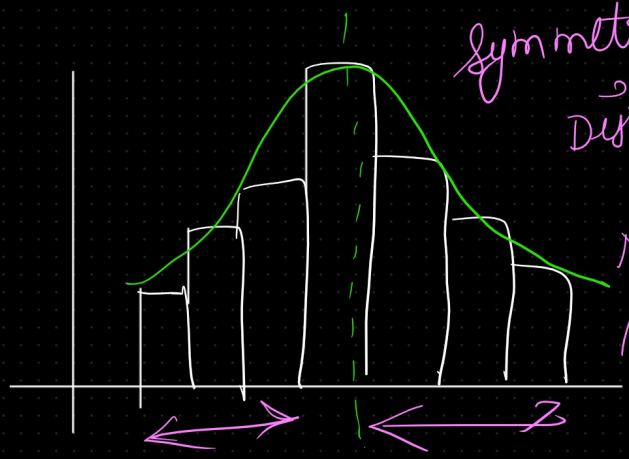


Histogram And Skewness

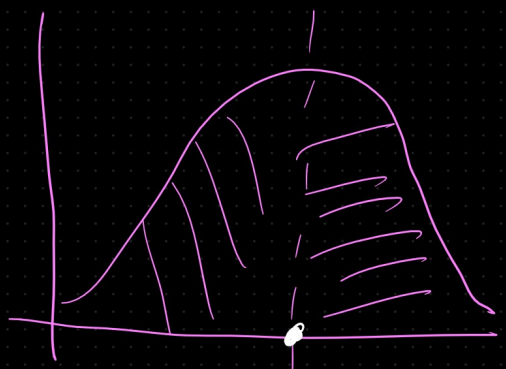
[Frequency]

data = {1, 1, 2, 3, 3, 3, 2, 2, 2, 4, 5, 6, 8, 16, 8}

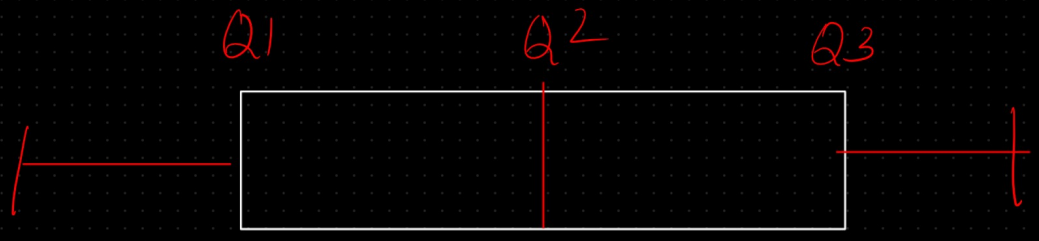




→ No Skewness



Mean, mode, median

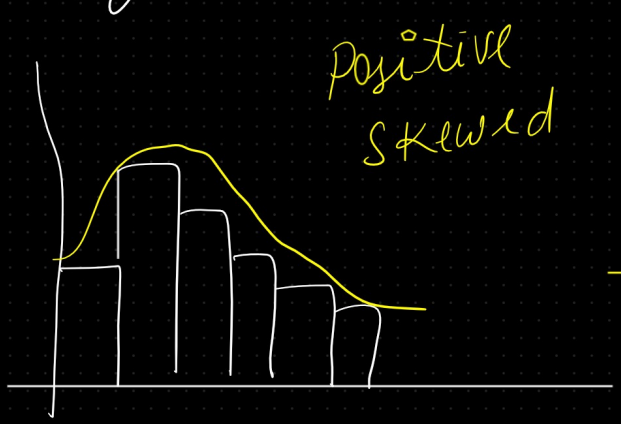


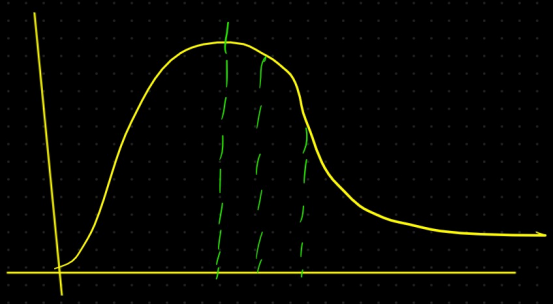
$Q2 - Q1 \approx Q3 - Q2$

mean = mode = median

② Right Skewed

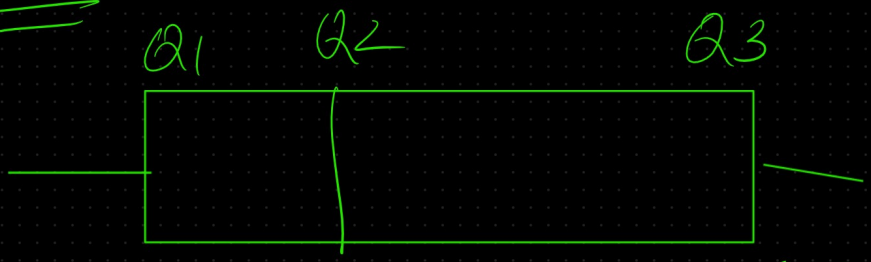
Positive Skewed





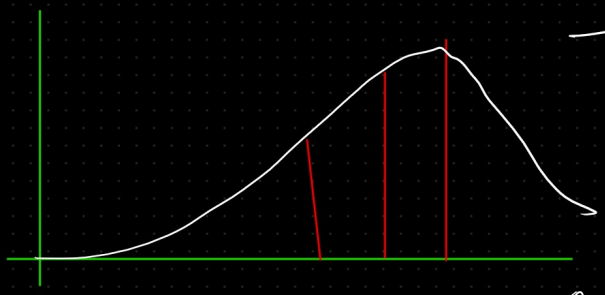
mean ≥ median ≥ mode

Box plot

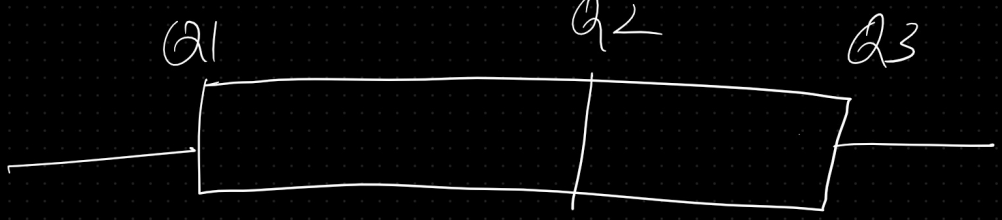


$Q3 - Q2 \geq Q2 - Q1$

③ Left Skewed Distribution



→ Negative Skewed



$Q2 - Q1 \geq Q3 - Q2$

mean ≤ median ≤ mode

Covariance and correlation

X	Y	[Relation ship b/w X and Y]
2	4	$X\uparrow$ $Y\uparrow$
5	6	$X\downarrow$ $Y\downarrow$
7	8	$X\downarrow$ $Y\uparrow$
10	12	$X\downarrow$ $Y\downarrow$

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

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

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

$= Cov(x,x) \Rightarrow$ Spread

$Cov(x, y)$ $\boxed{x\uparrow \; y\uparrow \quad x\downarrow \; y\downarrow}$ +ve Cov

$Cov(x, y)$ $\boxed{x\uparrow \; y\downarrow \quad x\downarrow \; y\uparrow}$ -ve Cov

Coefficient $[-1 \text{ to } 1]$

(1) Pearson correlation coeff.

$$\rho_{x,y} = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$$

a) $+1$ ← $Cov(x, y)$

b) -1 $Cov(x, y)$ →

∴ The more the value towards +1 the more +ve correlated is (x, y)

the more the values towards −1
the more −ve correlated (x, y)

(2) Spearman Rank Correlation
[−1 to 1]

$$\rho_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$