

Measures of central Tendency

* Mean

* Mode

* Median

→ Mean : The Average of the given numbers.

$$x = \{1, 2, 2, 3, 5, 7, 10\}$$

$$\text{Mean} = \frac{1+2+2+3+5+7+10}{7}$$
$$= 4.28$$

→ Median : The central element data after sorting all the values.

$$x = \{21, 30, 14, 17, 19, 10\}$$

Total 6 data points

* so median will be mean of central data.

$$\text{Median} = \frac{14+17}{2} \Rightarrow 15.5$$

$$x = \{ 21, 30, 45, 3, 10 \}$$

odd number of Elements

* so median is middle element i.e. 45

mode: It's the most occurring data point in the given data.

$$x = \{ 1, 2, 3, 4, 2, 2, 6, 6, 4, 4, 7, 9, 4, 4, 4 \}$$

In the above, 4 appeared 6 times

so, Mode is 4

Measures of Dispersion

Variance:

It is the Average of summation of squares of difference between each data point to the mean of the data points.

Population Variance(σ^2)

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

where,

x_i = Data points.

μ = Mean of population.

N = Total Data points.

sample variance(s^2)

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

where,

x_i = Data points

\bar{x} = Mean of sample

n = Total Data points

eg: Let's have Data as $\{1, 2, 3, 4, 5\}$. Find population and sample variance.

population

$$x_i \quad (x_i - \mu)^2$$

1	4
2	1
3	0
4	1
5	4

mean = $\frac{4}{5} = 0.8$
 μ

$$x_i \quad (x_i - \bar{x})^2$$

1	4
2	1
3	0
4	1
5	4

mean = $\frac{4}{5} = 0.8$
 \bar{x}

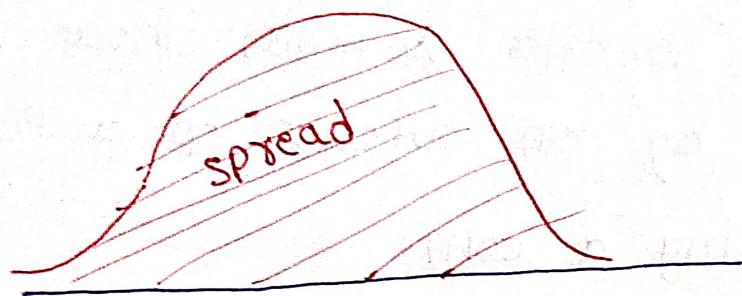
sample

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

$$= \frac{10}{5} \Rightarrow 2$$

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

$$= \frac{10}{4} \Rightarrow 2.5$$



- * Higher the variance, Higher the spread will be.
- * Lower the variance, Lower the spread will be.

standard Deviation:

- * simply, It's square root of variance.

Population

$$\text{If } \text{variance}(\sigma^2) = 100$$

$$\text{Then } \text{std}(\sigma) = 10$$

sample

$$\text{If } \text{variance}(s^2) = 100$$

$$\text{Then } \text{std}(s) = 10.$$

Random Variables:

- * It is a numeric variable, whose value is determined by an outcome of a Random event.

Eg: → Flipping a coin
→ Rolling a die

sets:

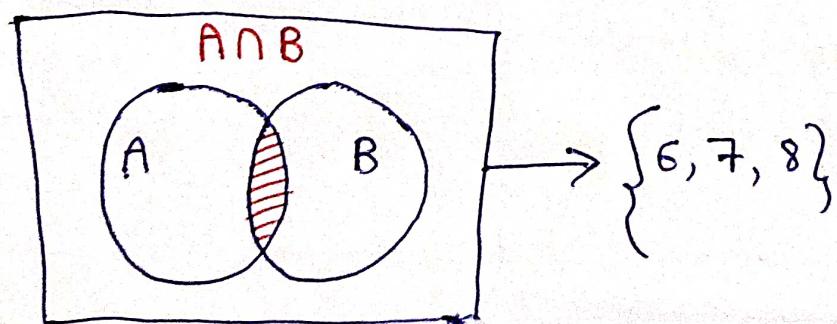
Let's consider 2 sets A & B

$$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

$$B = \{6, 7, 8\}$$

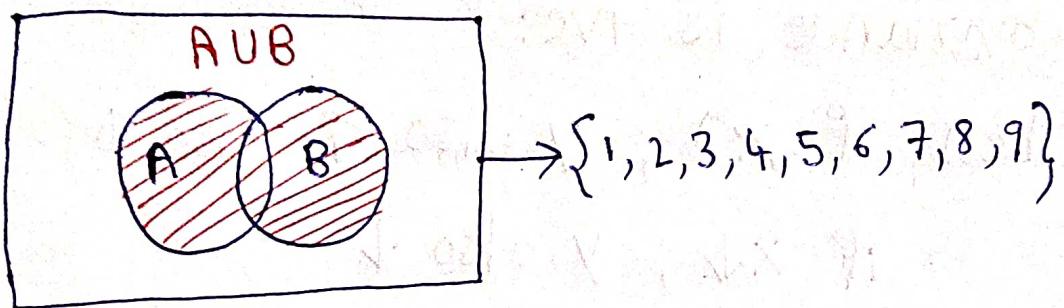
i) Intersection ($A \cap B$):

- * It's the common data on Both sets. $A \in B$.



2) union ($A \cup B$):

It's combination of all data points of A and B.



3) subset:

A is subset of B \rightarrow False

B is subset of A \rightarrow True

4) superset:

A is superset of B \rightarrow True

B is superset of A \rightarrow False

Covariance:

- * It is a metric to find how two variables are related to each other.

If covariance is +ve:

- Then if $x \uparrow$, $y \uparrow$ also ↑
if $x \downarrow$, $y \downarrow$

If covariance is -ve:

- Then, if $x \downarrow$, $y \uparrow$
if $x \uparrow$, $y \downarrow$

If covariance is zero:

- It represents there is no relation b/w x and y.

Advantage

Finds Relationship b/w
x and y

disadvantage

It doesn't have limit
values

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n-1}$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
2	3	-2	-2	4
4	5	0	0	0
6	7	2	2	4
$\bar{x} = 4$		$\bar{y} = 5$		

$$\Rightarrow \frac{4+0+4}{2} \Rightarrow \frac{8}{2} \Rightarrow 4 \rightarrow +ve$$

As it's +ve, if $x \uparrow$, then y also \uparrow
 if $x \downarrow$, then y also \downarrow

Pearson corelation:

It will give us corelation Ranges b/w -1 and 1

- * If values are +ve, then there is +ve corelation.
- * If -ve, then there is -ve corelation.
- * If 'zero', then there is no corelation.

$$\rho_{(x,y)} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

Spearman corelation:

Instead of x, y we will use $\text{Rank}(x), \text{Rank}(y)$ to calculate the pearson corelation for this.

$$\rho_s = \frac{\text{cov}(\text{R}(x), \text{R}(y))}{\sigma_{\text{R}(x)} \sigma_{\text{R}(y)}}$$

x	y	R(x)	R(y)
9	1	1	4
6	7	2	2
5	3	3	3
4	9	4	1

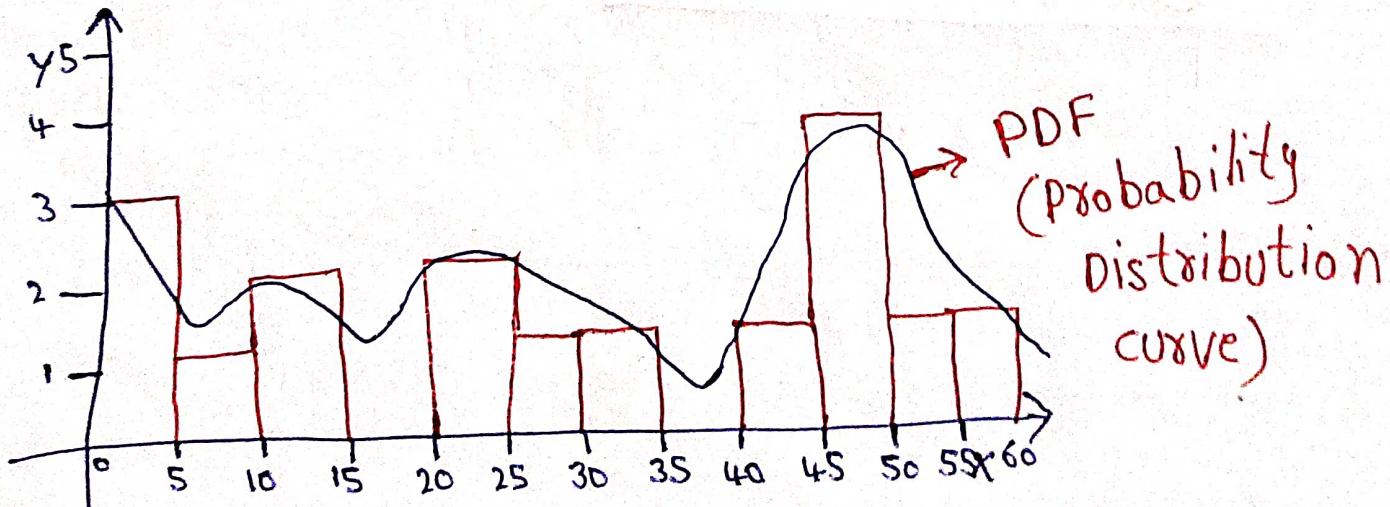
Histogram:

- * It's the representation of the spread of Data Points.
- * we have to draw a smoothening curve to get the probability Distribution curve

Eg: Let's take the data as

$$x = \{1, 2, 4, 9, 11, 14, 21, 23, 25, 32, 43, 45, 46, 47, 49, 51, 55\}$$

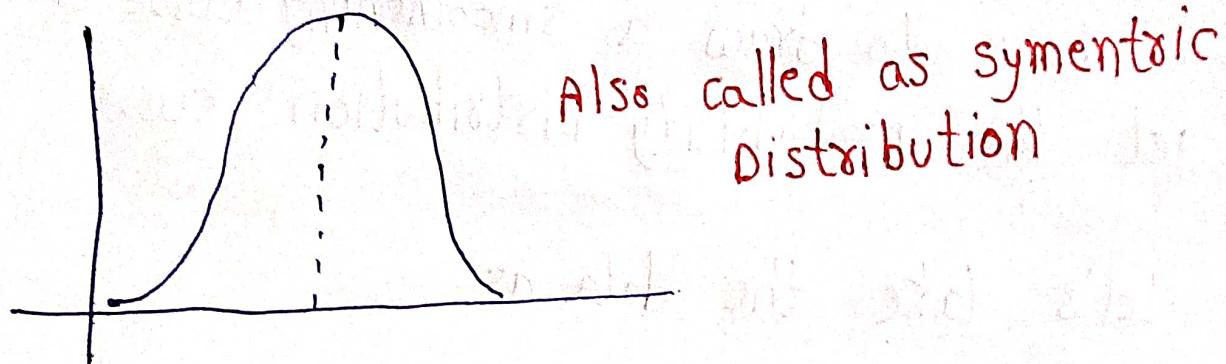
Let's take bin size = 5.



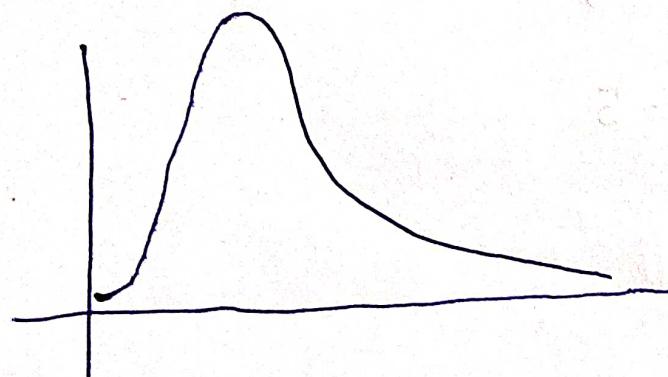
Types of Histogram PDF curves:

Below are some basic distributions.

Normal/Gaussian:



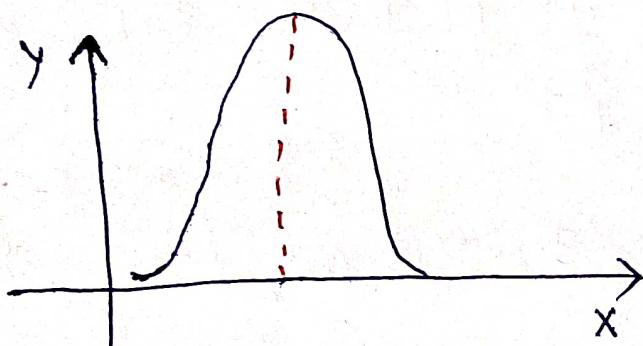
Log Normal:



Skewness:

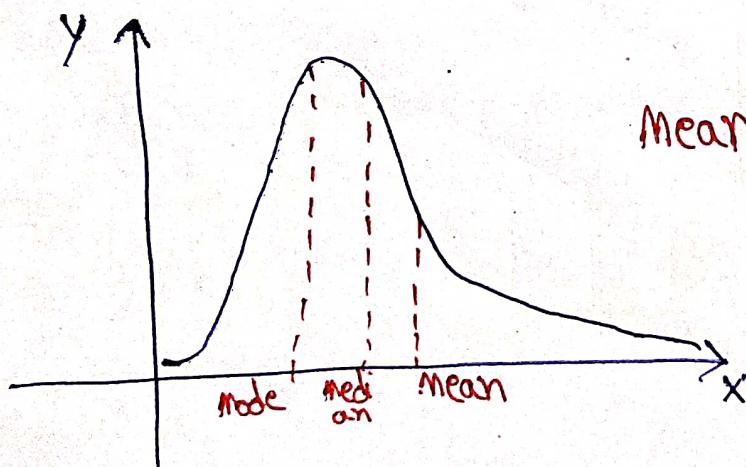
- * skewness is a metric to understand the type of data distribution.
- * we can make better statistical decisions by understanding it.

No skewness:



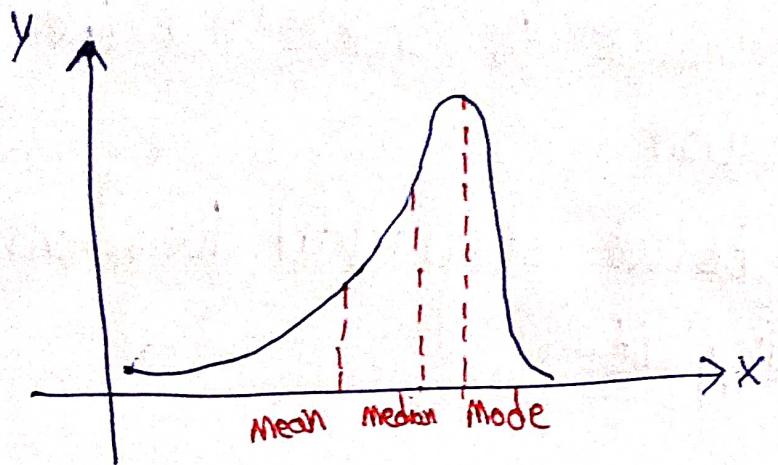
Here mean, median and Mode are almost similar.

+ve skewness/Right skewness:



Mean > Median > Mode

-ve skewness/Left skewness:



Mode > Median > Mean