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.

Eg: $X = \{21, 30, 14, 17, 19, 100\}$

* sort the values

$X = \{14, 17, (19, 21), 30, 100\}$

Total 6 Data points, i.e **even** number

* so median will be mean of central Data.

$$\text{Median} = \frac{19+21}{2} \Rightarrow \boxed{20}$$

Eg: $X = \{21, 30, 45, 3, 10\}$

sorted → $X = \{03, 10, (21), 30, 45\}$

odd no. of. Data points.

* so median is center element, i.e $\boxed{21}$

Mode: It's the most occuring 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: Lets have Data as $\{1, 2, 3, 4, 5\}$. Find population and sample variance.

Population

x_i	$(x_i - \mu)^2$
1	4
2	1
3	0
4	1
5	4
Mean $= 3$ (μ)	Sum $= 10$

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

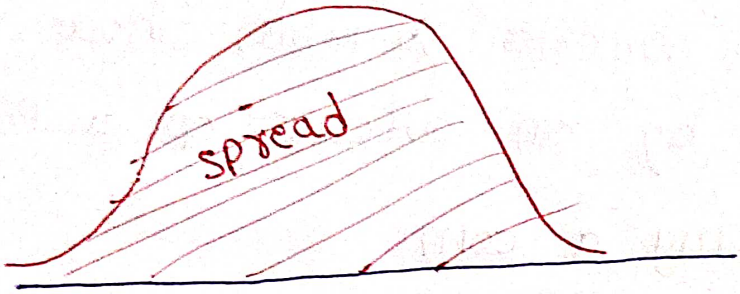
$$= \frac{10}{5} \Rightarrow \boxed{2}$$

Sample

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

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

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



* 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	sample
If variance(σ^2) = 100	If Variance (s^2) = 100
Then std(σ) = 10	Then 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:

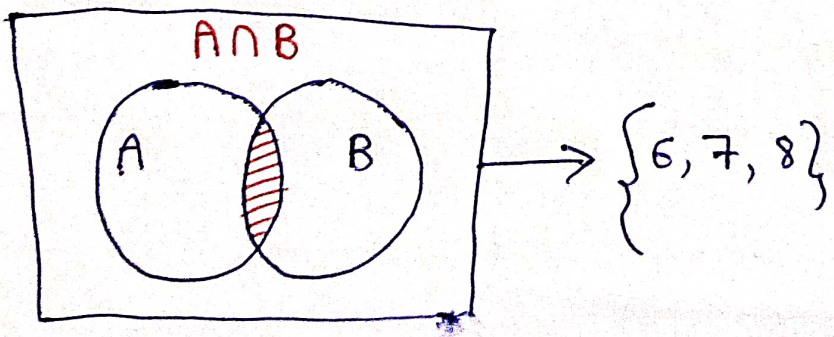
Let's consider 2 sets A & B

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

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

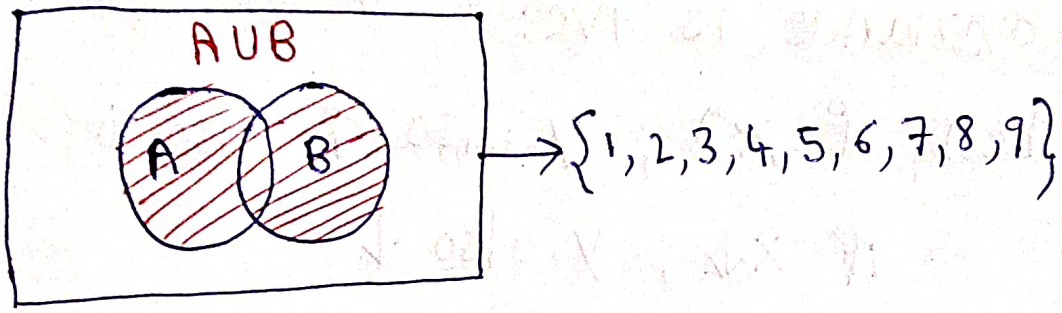
i) Intersection ($A \cap B$):

* It's the common data on Both sets. A & B.



2) union (A U B):

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



3) subset:

A is subset of B → False

B is subset of A → True

4) superset:

A is superset of B → True

B is superset of A → False

Covariance:

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

If covariance is +ve:

→ Then if x↑, y ~~also ↑~~ also ↑

if x↓, y also ↓

If covariance is -ve:

→ Then, if x↓, y ↑

if x↑, Then y↓

If covariance is zero:

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

Advantage	Disadvantage
Finds Relationship b/w X and Y	It doesn't have Limit values

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

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↑, then y also ↑
if x↓, then y also ↓

Pearson coreleation:

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

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

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

Spearman coreleation:

Instead of x, y we will use Rank(x), Rank(y) to the pearson coreleation for this.

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

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

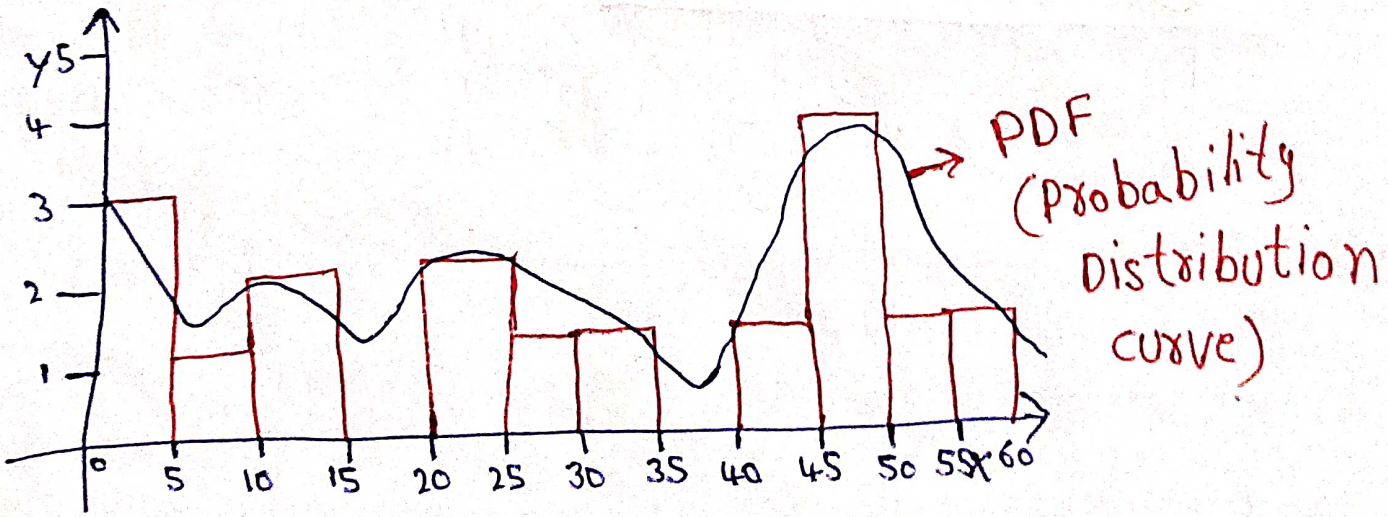
Histogram:

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

eg: Let's take the data as

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

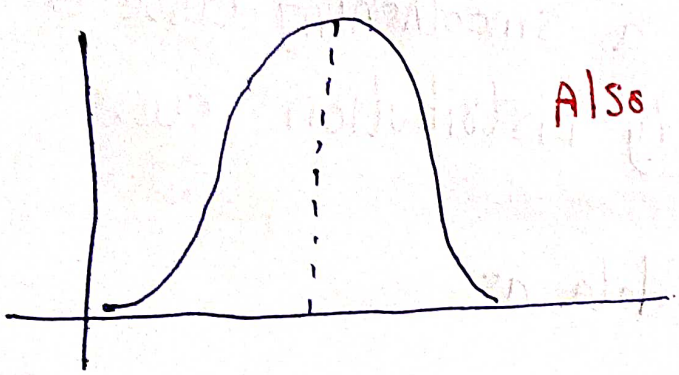
Lets Take bin size = 5



Types of Histogram PDF curves:

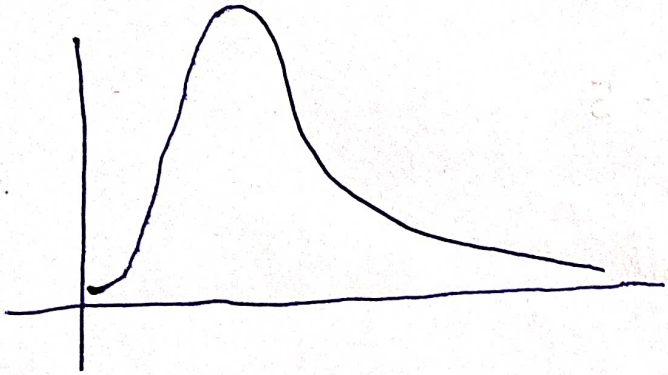
Below are some basic Distributions.

Normal/Guassian:



Also called as symentric Distribution

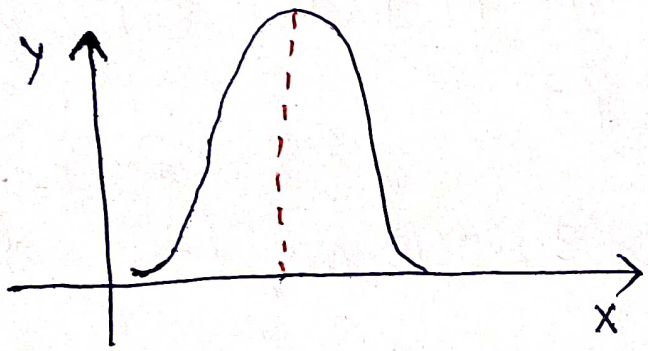
Log Normal:



Skewness:

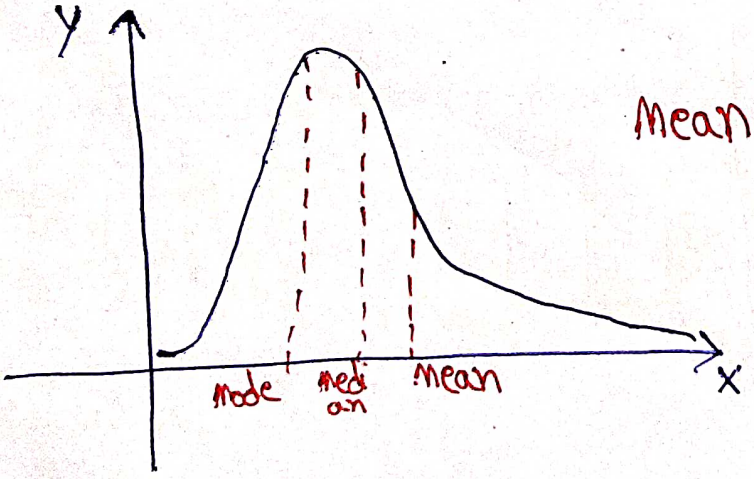
* skewness is a metric to understand the type of data distribution.
* we can make better statistical decissions by understanding it.

No skewness:



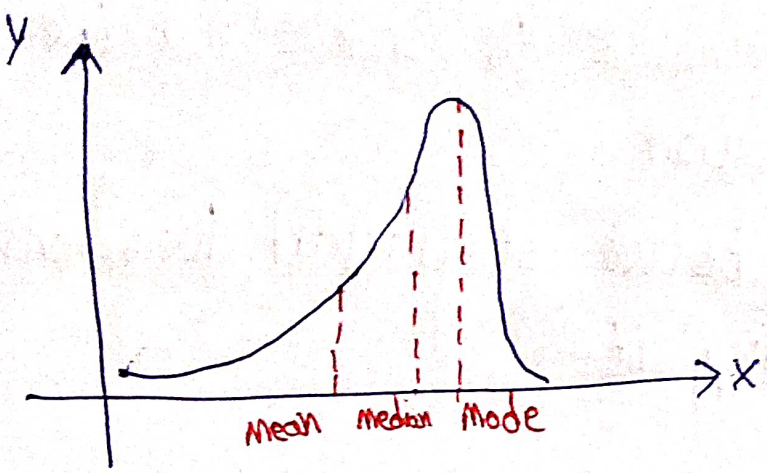
Here mean, median and Mode are almost similar.

+ve skewness / Right skewness:



Mean > Median > Mode

-ve skewness/Left skewness:



Mode > Median > Mean