

## statistics

\* It is the science of collection, organizing and analyzing data.

what is Data:

Data is set of facts or Information.

e.g :

A Bank has opened an ATM at a Location A and there is another Location B.

The Bank want to know whether it's useful to open a Bank at B which is 40 km from A.

\* For the above example we can use statistics to help Bank people to execute better decision.

## Types of statistics:

### ① Descriptive :

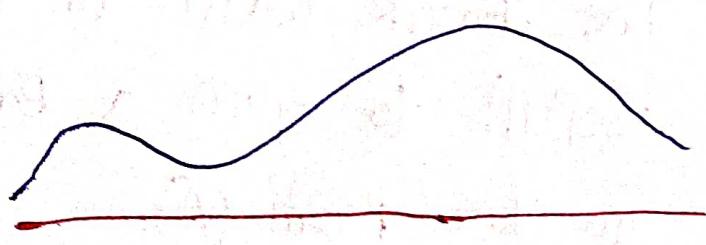
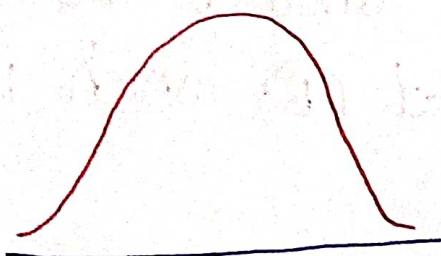
\* It consists of organising and summarizing the data.

→ Applying different operations to take new data.

e.g.: Measure of central Tendency

Measure of Dispersion.

Histograms, Bar chart, pie



### ② Inferrential:

using data you have measured to form conclusion.

e.g.: z-test.

t-test.

## sample questions:

- \* Let's say there are 50 students in a maths class in the university, and we've collected the height of the student in the class.  
[175, 180, 160, 140, 130, 140, 140, ...]

### Descriptive:

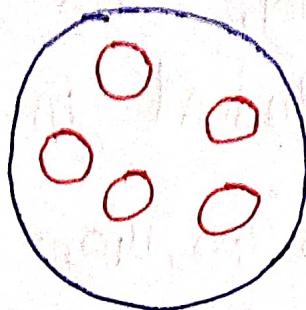
- \* what is the average height.
- \* what is the most common height.

### Infrential:

- \* Are the height of the students in the classroom similar to what you expect in entire college.

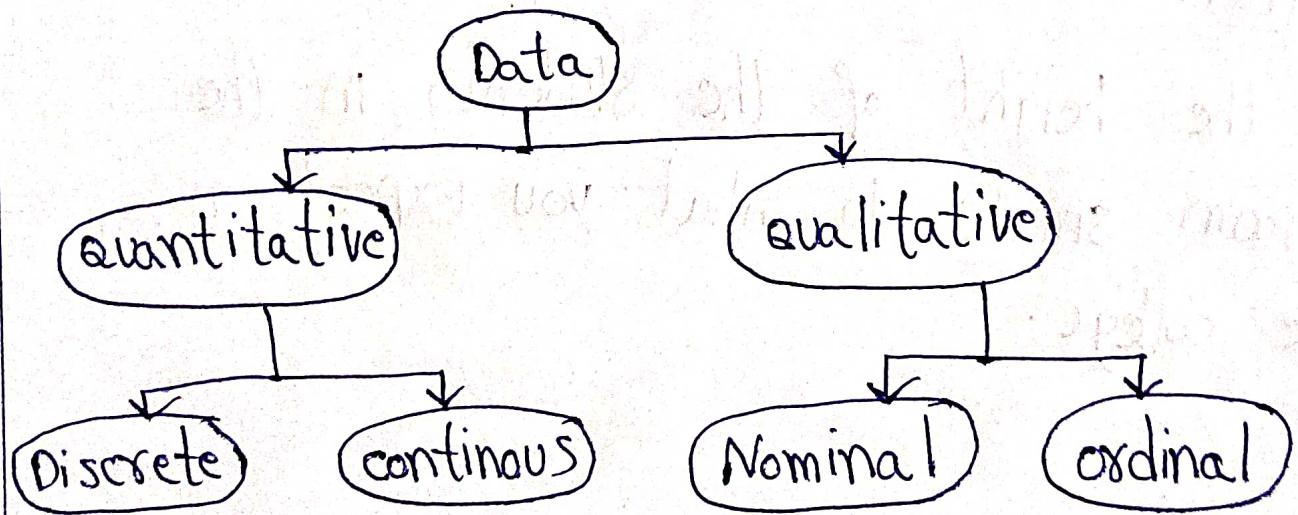
## Sample and population Data:

e.g.: Exit poll



Survey Agency will take sample data and makes conclusion of entire data.

## Types of Data:



quantitative: numerical

discrete: Indicates specific whole numbers.

e.g.: No. of Bank Accounts.

No. of children in a family.

continuous: May have any value

e.g.: Heights, weights, Temperature, speed

qualitative: categorical.

Nominal: Fixed set of categories

e.g.: Gender, Blood group, color

ordinal: which can have Ranks.

e.g.: Best, Better, worse, worst.

## Scale of Measurement of Data:

### ① Nominal scale Data:

\* qualitative / categorical.

\* eg: Gender, Labels.

\* order doesn't matter.

eg:

Red → 5 → 50%

Blue → 2 → 20%

yellow → 3 → 30%

### ② ordinal scale Data:

\* Ranking and order matter.

\* Difference cannot be measured.

eg: Inter → 3

MS → 1

B.Tech → 2

10th → 4

### ③ Interval scale data:

- \* Rank and order matters
- \* Difference can be measured (Excluding ratio)
- \* Does not have "0" starting value.

e.g.: - Temperature.

30 F

60 F

80 F

90 F

### ④ Ratio scaled Data:

- \* order and rank matter.
- \* Difference & Ratio measurable.
- \* can have "0" starting value.

e.g.: Marks      100, 70, 65, 85, 90, 72

## 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.

$$\text{median} = \frac{19+21}{2} \rightarrow 20$$

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

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

odd no. of. Data points.

\* so median is center element, i.e 21

**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( $\sigma^2$ )

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

where,

$x_i$  = Data points.

$\mu$  = 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 =  $\frac{4}{5} = 0.8$    sum = 10

### sample

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

1	4
2	1
3	0
4	1
5	4

$\bar{x} = \frac{4}{5} = 0.8$    sum = 10

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

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

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

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



- \* 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:

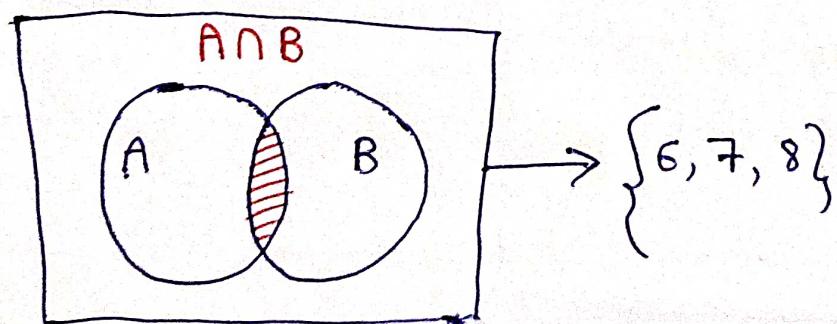
Let's consider 2 sets A & B

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

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

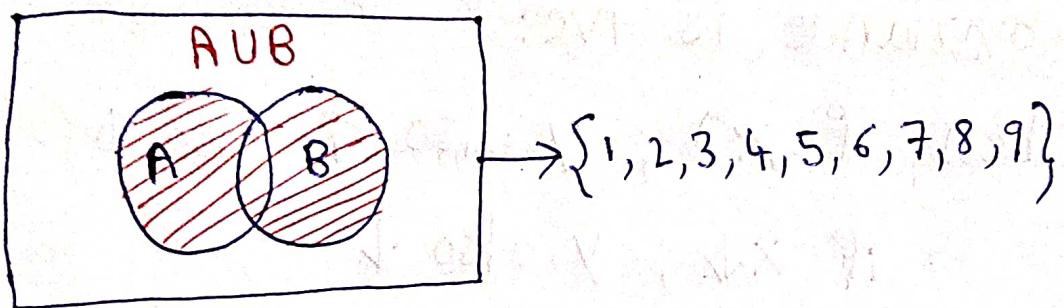
## i) Intersection ( $A \cap B$ ):

- \* 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

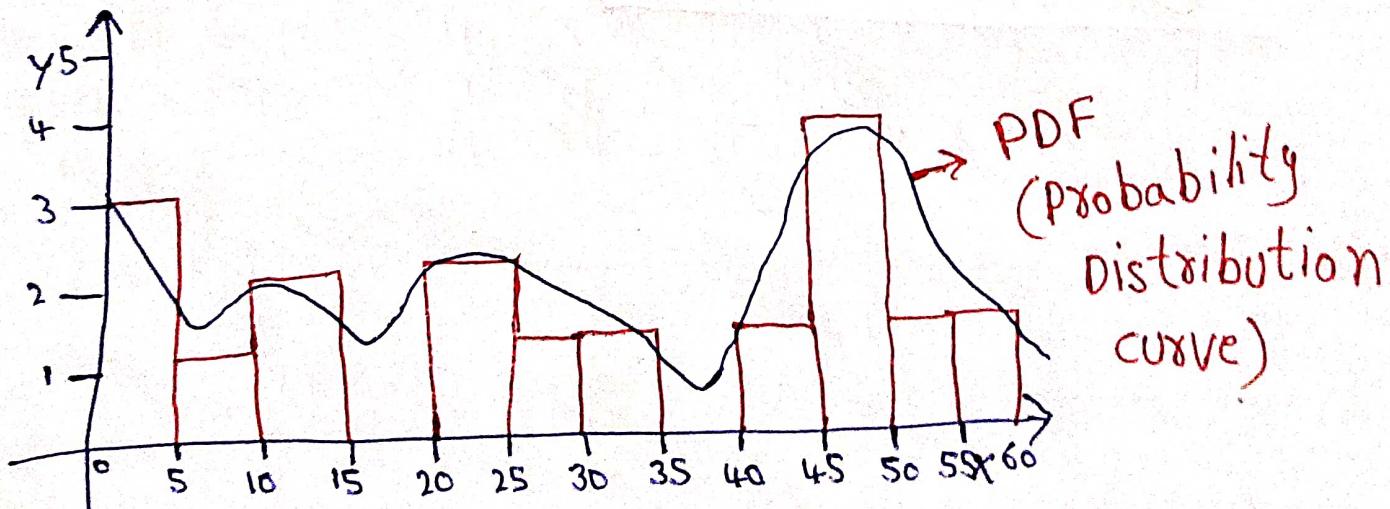
## Histogram:

- \* 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

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

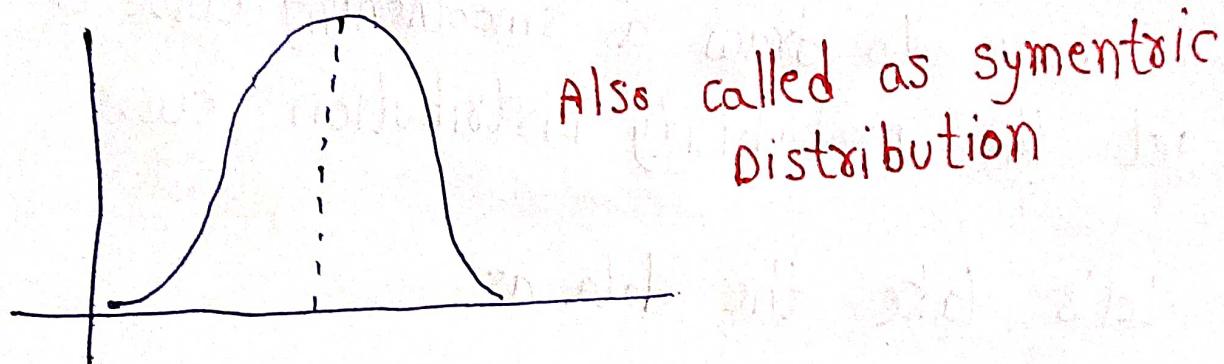
Let's take bin size = 5.



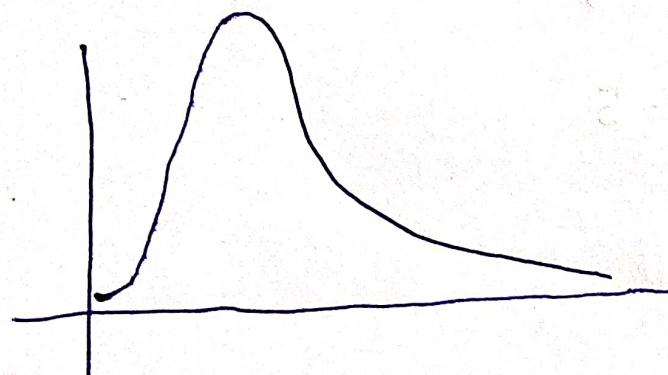
Types of Histogram PDF curves:

Below are some basic distributions.

Normal/Gaussian:



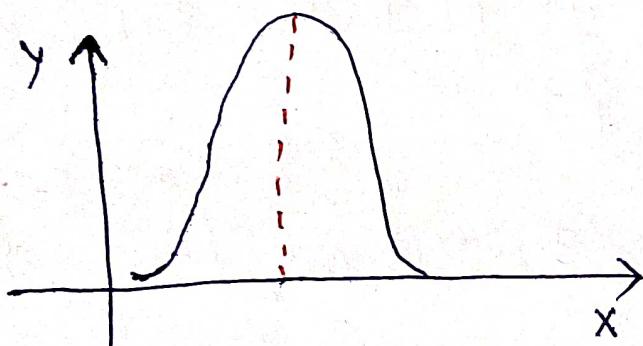
Log Normal:



## Skewness:

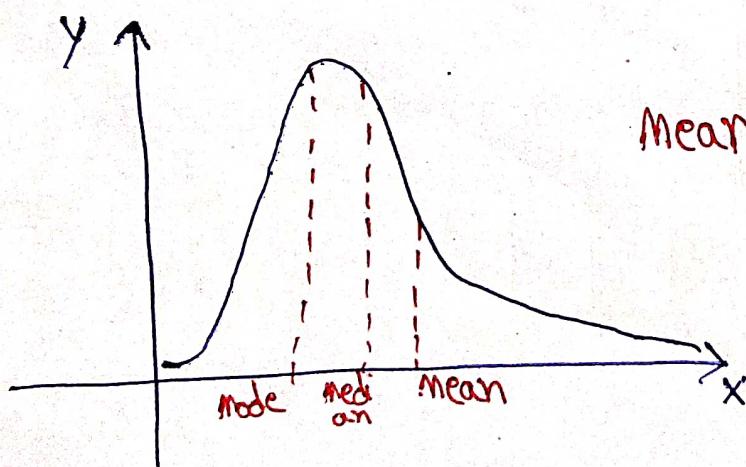
- \* skewness is a metric to understand the type of data distribution.
- \* we can make better statistical decisions by understanding it.

## No skewness:



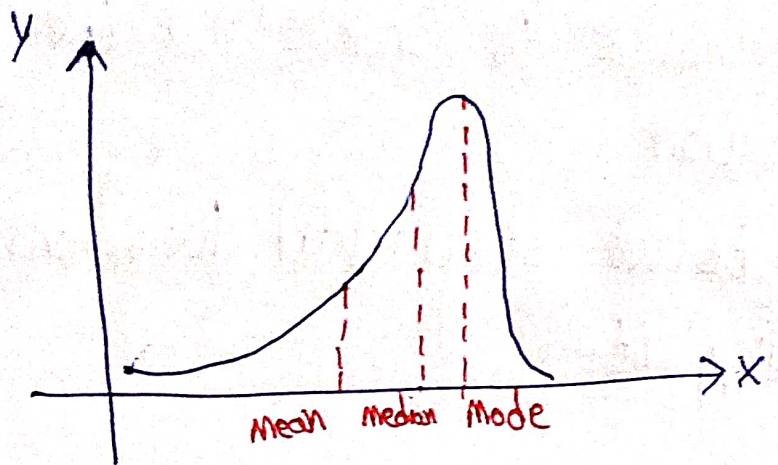
Here mean, median and Mode are almost similar.

## +ve skewness/Right skewness:



Mean > Median > Mode

-ve skewness/Left skewness:



Mode > Median > Mean

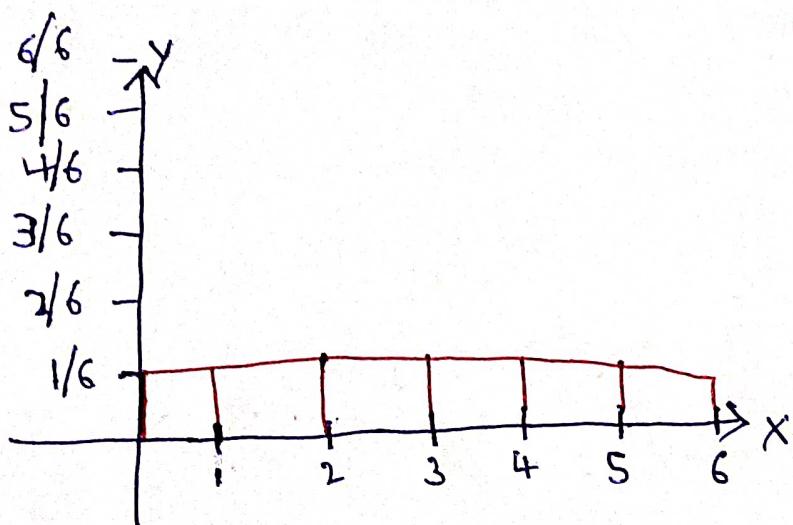
## Probability Mass Function:

\* used to Represent Discrete Random Variables

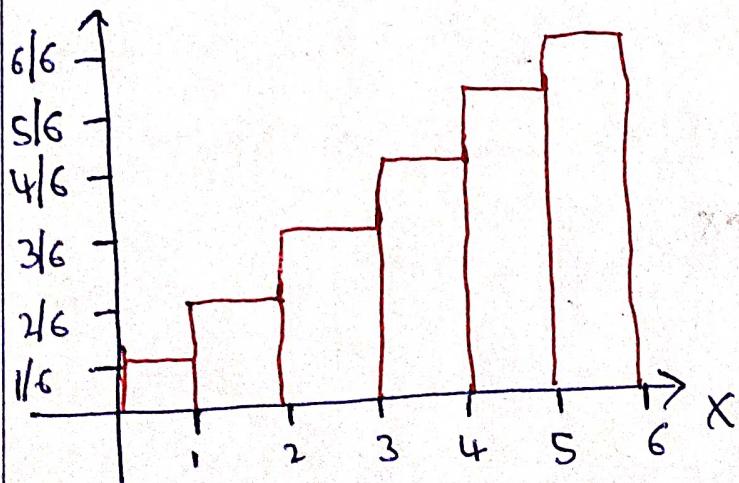
For eg: Rolling a Dice

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

$$P_X(1) = 1/6, P_X(2) = 1/6, \dots$$



## cumulative probability :

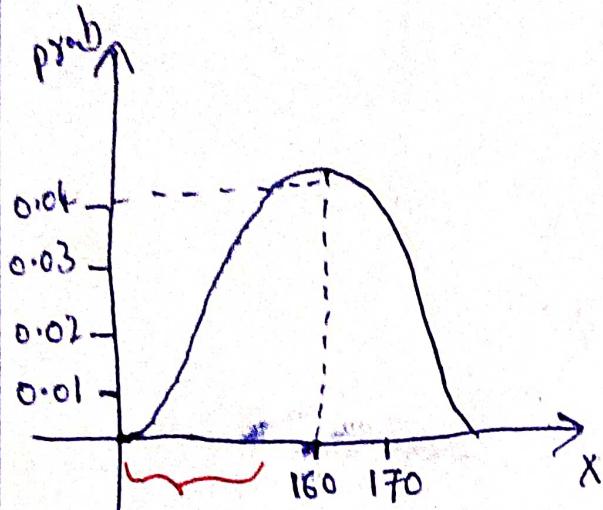


\* It's the cumulative sum of all the Data points.

## Probability Density Function:

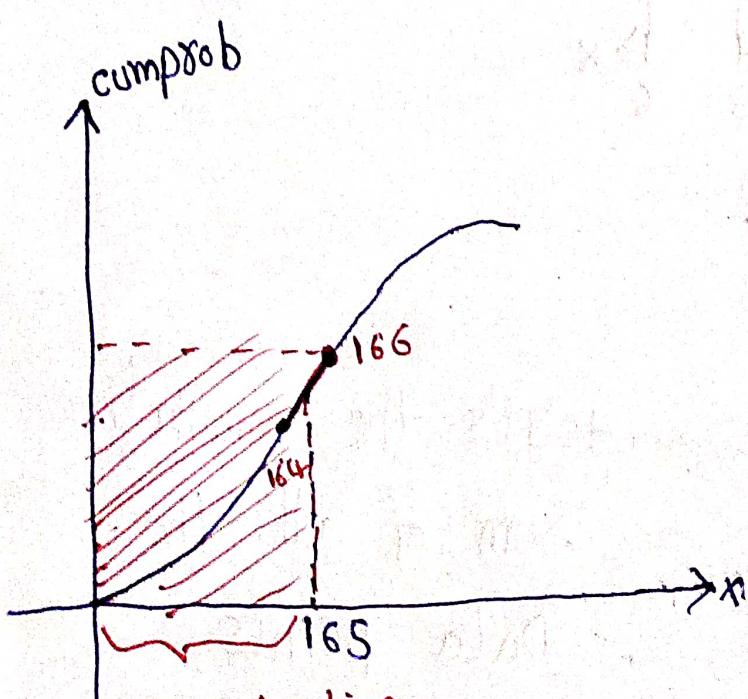
\* used to represent continuous Random variables.

Eg: Heights



50% of  
Entire Distribution

CDF:



50% of Entire  
distribution.

## Bernoulli Distribution:

- \* It is the distribution where two discrete outcomes will be represented by  $P$  and  $1-P$
- eg: Tossing a coin

$$P(H) = P = 0.50$$

$$\Rightarrow P(T) = 1 - P \rightarrow q$$

mean of Bernoulli Distribution :  $P$

Variance :  $Pq$

std deviation :  $\sqrt{Pq}$

## Binomial Distribution:

- \* It is same as Bernoulli, But two Discrete outcomes will be done for  $n$  number of Times.

eg: Toss a coin 13 Times

Mean :  $NP$

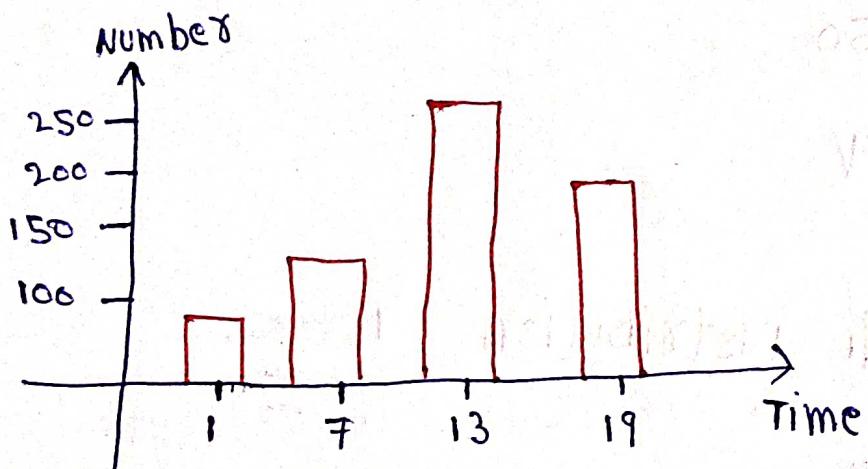
Variance :  $NPq$

std Deviation :  $\sqrt{NPq}$

## Poisson Distribution:

\* Describes the number of events occurring at a certain period of time.

e.g.: No of people visiting hospital every  $\boxed{6}$  hours.



$$\lambda = 135$$

\* (1) Represents Expected no. of events at every Time Interval

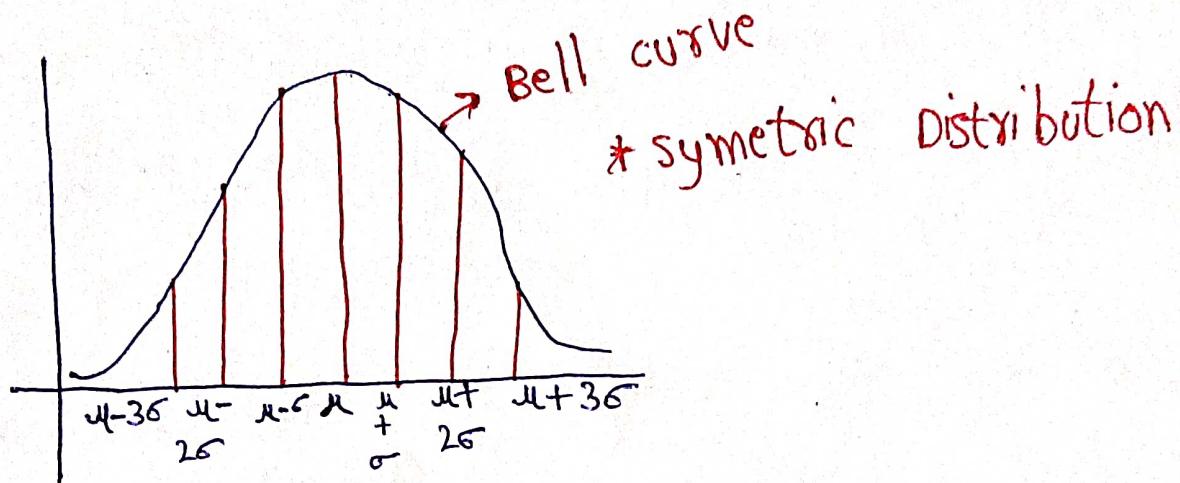
$$\text{mean} \rightarrow \lambda t$$

$$\text{Variance} \rightarrow \lambda t$$

\*  $t$  = time Interval

$\lambda$  = Expected no. of Events

## Normal ( $\sigma$ ) Gaussian Distribution:



## Empirical Rule :

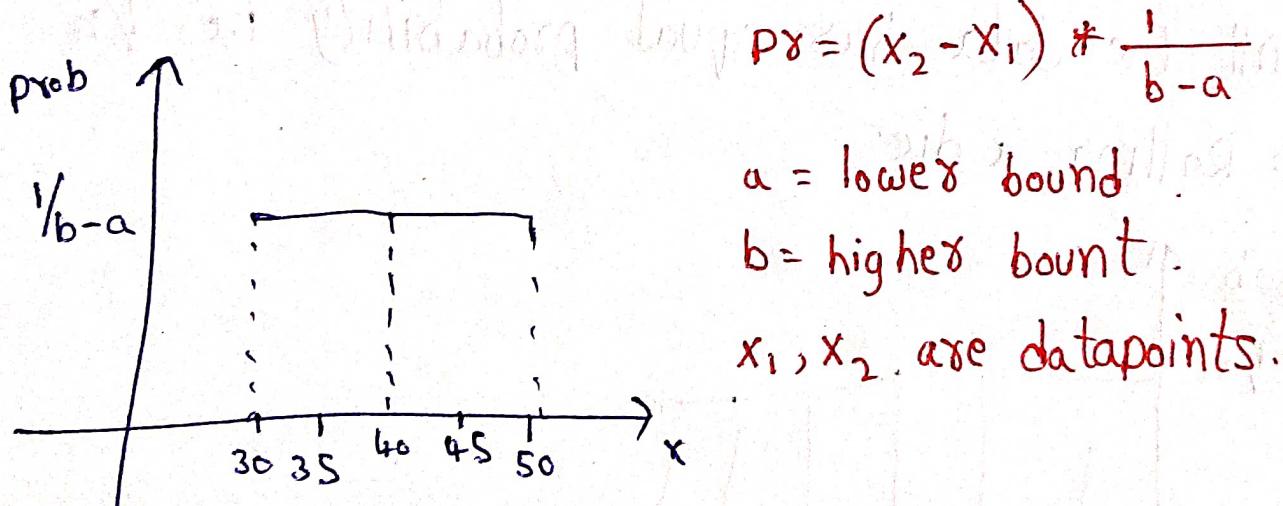
- \* 68% Data lies b/w  $\mu - \sigma$  and  $\mu + \sigma$
- \* 95% Data lies b/w  $\mu - 2\sigma$  and  $\mu + 2\sigma$
- \* 99.7% Data lies b/w  $\mu - 3\sigma$  and  $\mu + 3\sigma$

e.g.: Based on so much of Analysis, Researchers Found Height, weight, iris, etc... will mostly have gaussian Distribution

## uniform distribution

### continuous uniform distribution:

- \* In uniform distribution, the probability of getting outcomes is equal
- \* In continuous uniform distribution, In between a specified Range the probability of getting outcomes is same.



$$\begin{aligned} Pr(x \text{ b/w } 40 \text{ & } 45) &= (45 - 40) * \frac{1}{50 - 30} \\ &= 5/20 \\ &= 1/4 \rightarrow 0.25 \rightarrow 25\% \end{aligned}$$



$$P(X \text{ b/w } 30 \text{ & } 35) = (35-30) * \frac{1}{20}$$
$$= 0.25 \rightarrow 25\%$$

$$P(X \text{ b/w } 30 \text{ & } 45) = (45-30) * \frac{1}{20}$$

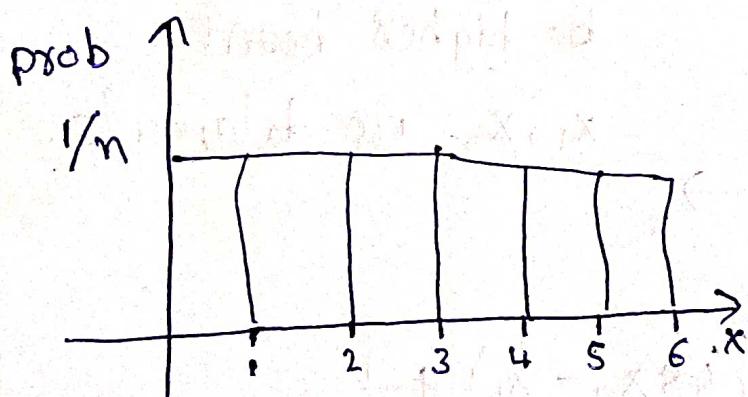
Mean :  $a+b/2 = 15/20$

Variance :  $(b-a)^2/12 = 0.75 \rightarrow 75\%$

### Discrete uniform distribution:

\* All the data has equal probability i.e.,  $1/n$

e.g: Rolling a die

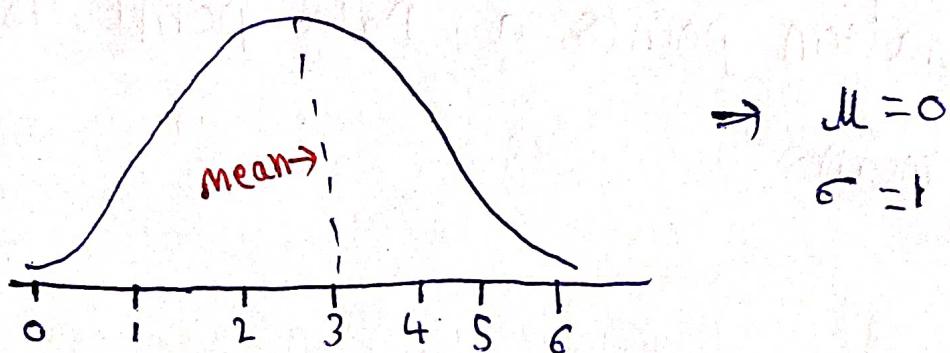


$$P(X=1) = 1/n \rightarrow 1/6$$

## standard Normal distribution:

\* It is the process of converting Normal Distribution into -ve, zero and +ve Data

For eg this is Normal distribution:



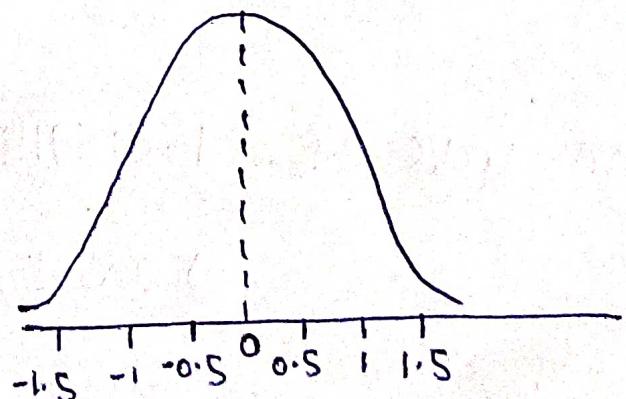
Let's Apply z score on all Data

$$z = \frac{x_i - \mu}{\sigma}$$

$$\mu = \text{mean} = 3$$

$$\sigma = \text{std Deviation} = 2$$

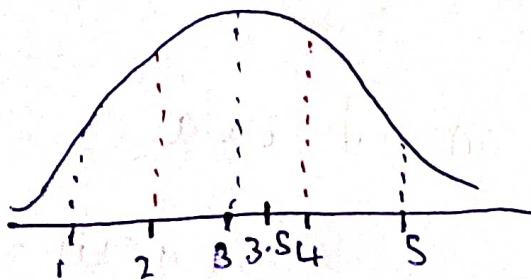
data	z-score
0	-1.5
1	-1
2	-0.5
3	0
4	0.5
5	1
6	1.5



## z-scores :

- \* z-score is used to standardize the data.
- \* 
$$z = \frac{x_i - \mu}{\sigma}$$
- \* we can also relate 'z' with how many std deviation points a data point is away from mean.

e.g.:



\* Find px of Region above 3.5  $\Rightarrow$

$$z = \frac{3.5 - 3}{1} \rightarrow 0.5 \rightarrow \text{Find its value in } z\text{-Table}$$

$0.6915 \quad \left. \right\} \text{ Its the region out of } 1 \text{ Before } 3.5$

$$\rightarrow \text{Region above } 3.5 = 1 - 0.6915 \Rightarrow 0.3085 \\ \Rightarrow 30.85\%$$

chi square:

\* It is a method used to check how efficient are 2 sets of data.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = observed data

E = expected data

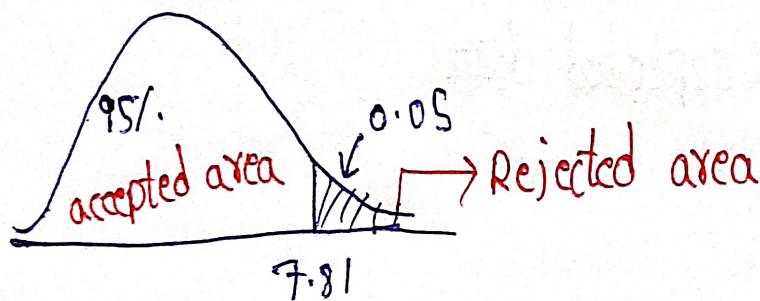
e.g. :

	Expected	observed	$(O - E)^2$	$(O - E)^2/E$
n=4	125	150	625	5
	125	200	5625	45
	125	100	625	5
	125	50	5625	45
				100

$$\alpha = 0.05 \text{ (given)}$$

$$\begin{aligned}\text{degree of freedom} &= n - 1 \\ &= 4 - 1 \\ &= 3\end{aligned}$$

For dof 3,  $\alpha = 0.05$ , From chi square table,  
critical value = 7.81



As we got chi square value we got is greater than 7.81 i.e.,  $100 > 7.81$ , we can consider Expected data is significantly different from observed data.

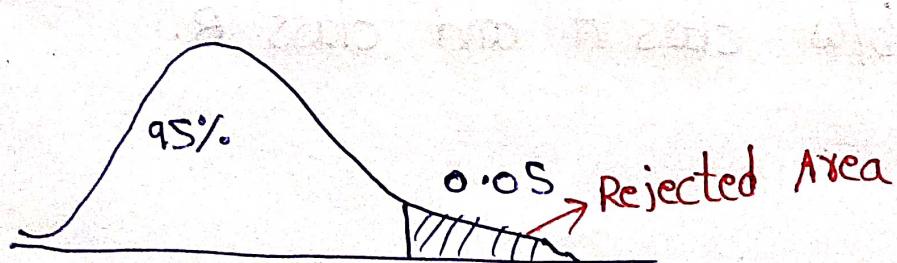
## F Test :

- \* It is used to find the comparison of variances of two data sets

$$F = \frac{\sigma(d_1)^2}{\sigma(d_2)^2}$$

e.g :

class	size	Variance	$\alpha = 0.05$
A	20	2.5	
B	20	1.8	



degree of freedoms:

$$df_1 = n - 1 = 20 - 1 = 19$$

$$df_2 = n - 1 = 20 - 1 = 19$$

At  $\alpha=0.05$  and  $df_1=19$ ,  $df_2=19$ , From F-Test table, critical value is 2.20

$$F = \frac{2.5}{1.8} \Rightarrow 1.39$$

$\therefore$  AS F value (1.39) < critical value (2.20)

we can say there is no significant difference b/w class A and class B.