statistics

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

what is Data:

Data is set of facts or Information.

eg:

A Bank has opened an ATM at a Location (A) and there is another Location (B). The Bank want to know whether it's usefull 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 decession.

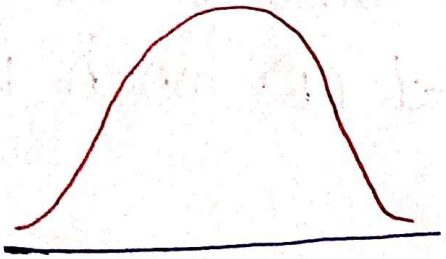
Types of statistics:

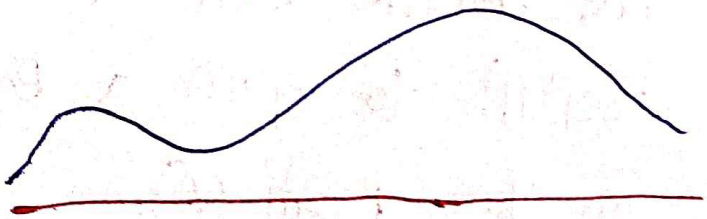
① Descriptive:

* It consists of organising and summarizing the data.

→ Applying different operations to take new data.

eg: Measure of central Tendency
Measure of Dispersion.
Histograms, Bar chart, pie





② Inferential:

using data you have measured to form conclusion.

eg: 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:

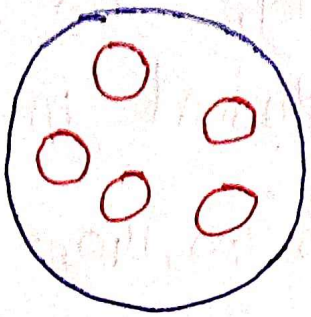
* what is the average height.
* what is the most common height.

Inferential:

* Are the height of the students in the classroom similar to what you EXPECT in Entire college.

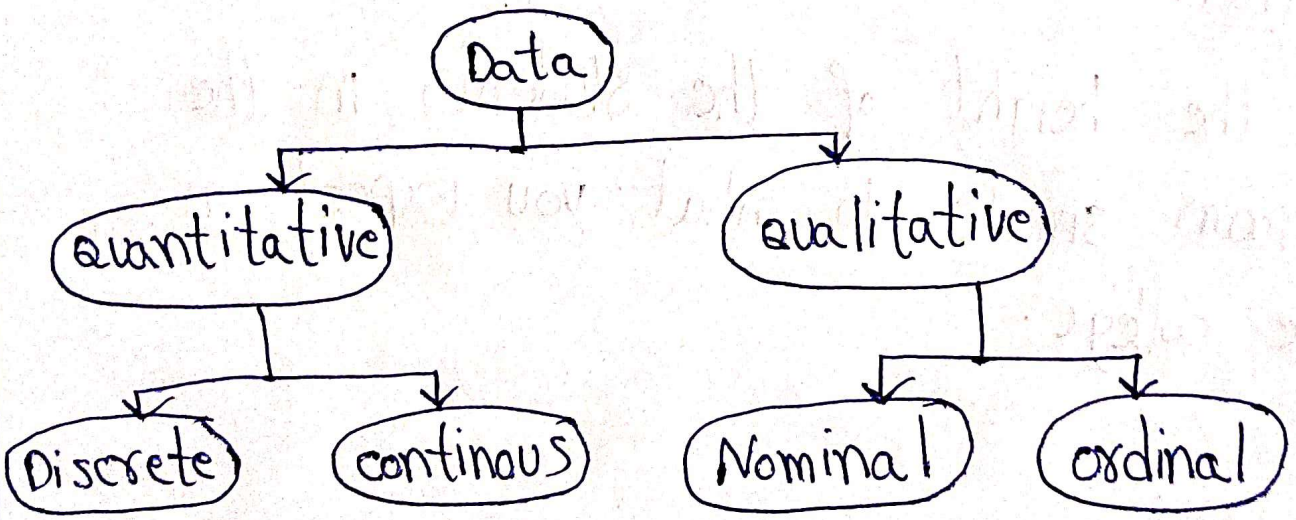
Sample and Population Data:

eg: Exit poll



survey Agency will take sample data and Makes conclusion of Entire data.

Types of Data:



Quantitative: Numerical

Discrete: Indicates specfic whole numbers.

eg: No. of. Bank Accounts.

No. of. children in a family.

continous: May have any value

eg: Heights, weights, Temperature, speed

Qualitative: categorical.

Nominal: Fixed set of categories

eg: Gender, Blood group, color

ordinal: which can have Ranks.

eg: 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.

Eg :- 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.

Eg: 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 → $X = \{03, 10, (21), 30, 45\}$

odd no. of. Data points.

* so median is center element, i.e 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 $= 10$

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

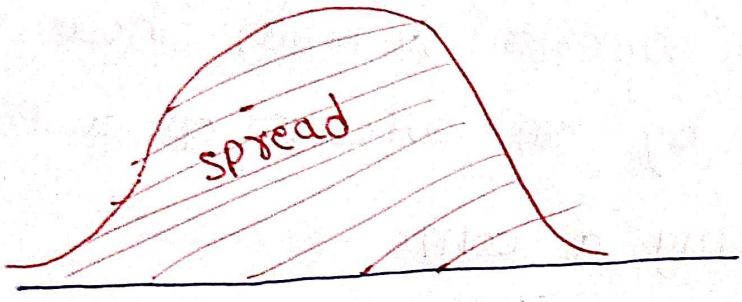
$$= \frac{10}{5} \Rightarrow \boxed{2}$$

Sample

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

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

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



* 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:

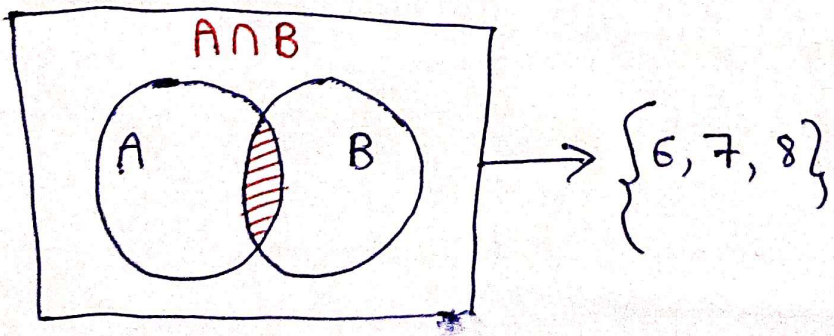
Let's consider 2 sets A & B

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

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

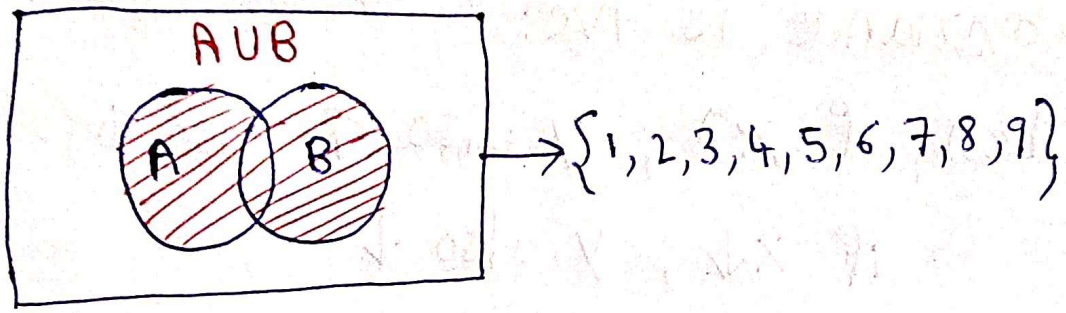
1) Intersection ($A \cap B$):

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



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

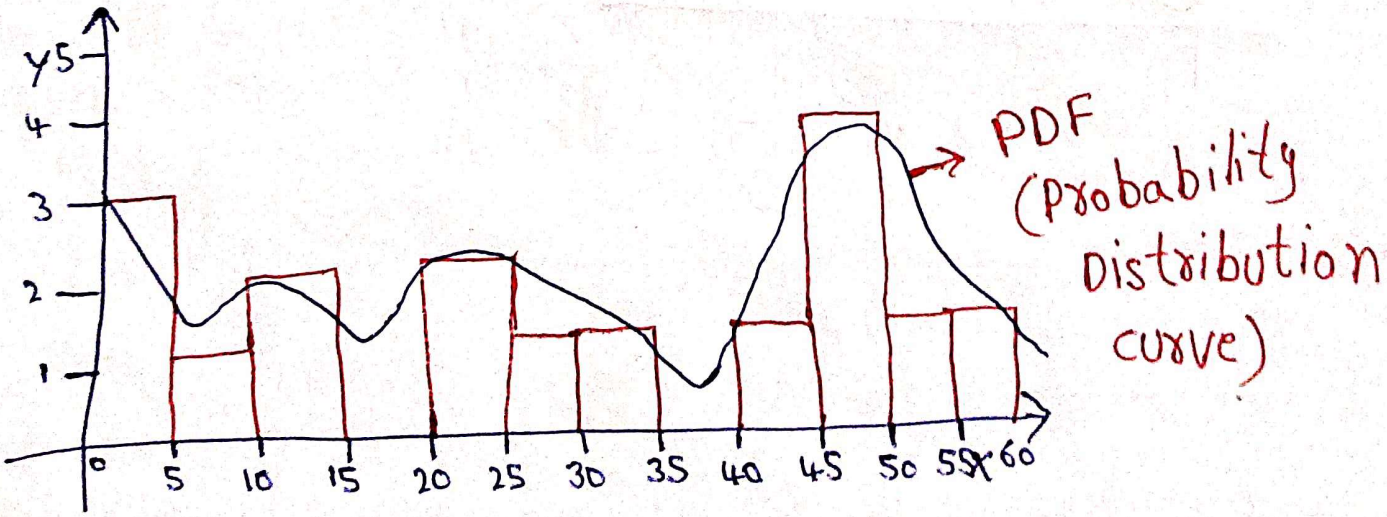
Histogram:

* 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

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

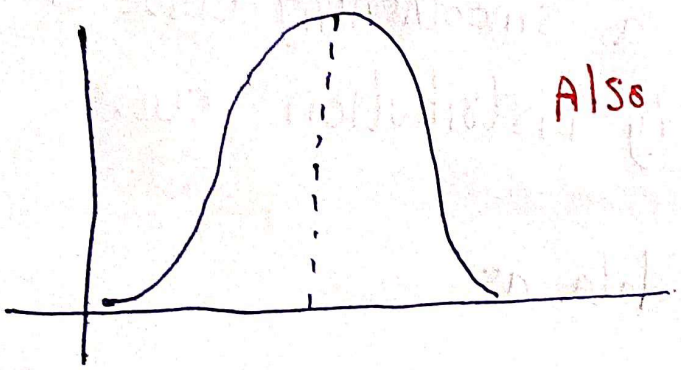
Lets Take bin size = 5



Types of Histogram PDF curves:

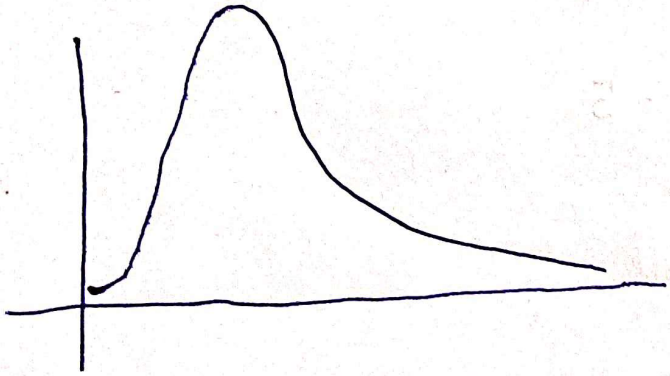
Below are some basic Distributions.

Normal/Guassian:



Also called as symentric Distribution

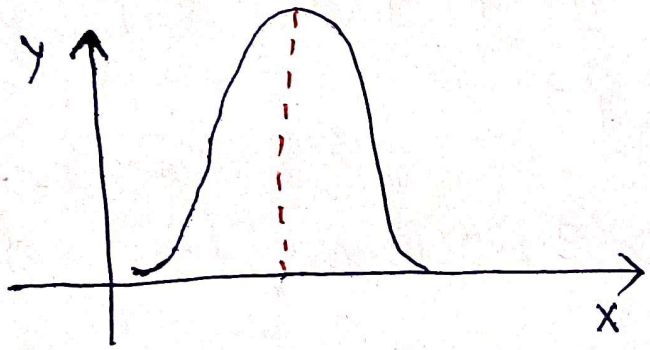
Log Normal:



Skewness:

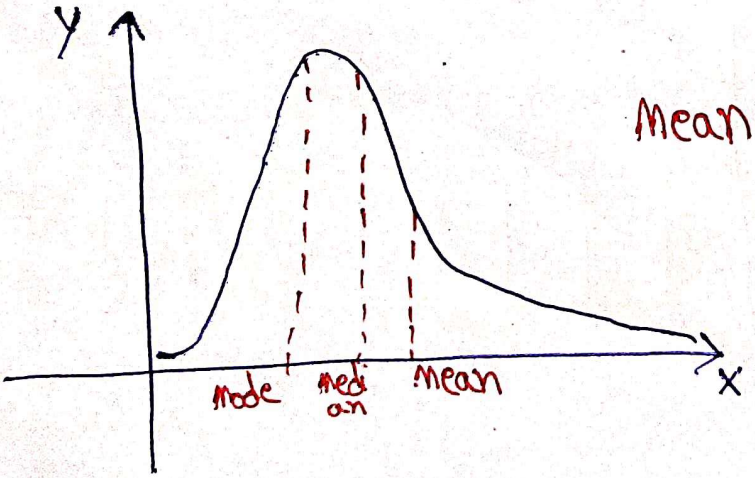
* skewness is a metric to understand the type of data distribution.
* we can make better statistical decissions by understanding it.

No skewness:



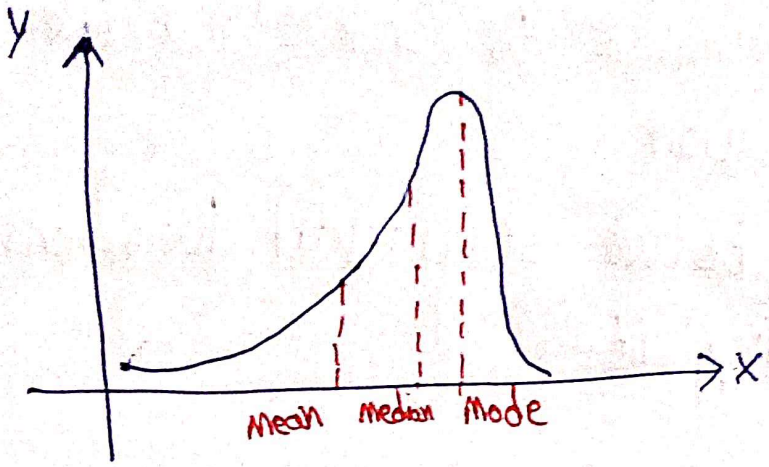
Here mean, median and Mode are almost similar.

+ve skewness / Right skewness:



Mean > Median > Mode

-ve skewness/Left skewness:



Mode > Median > Mean

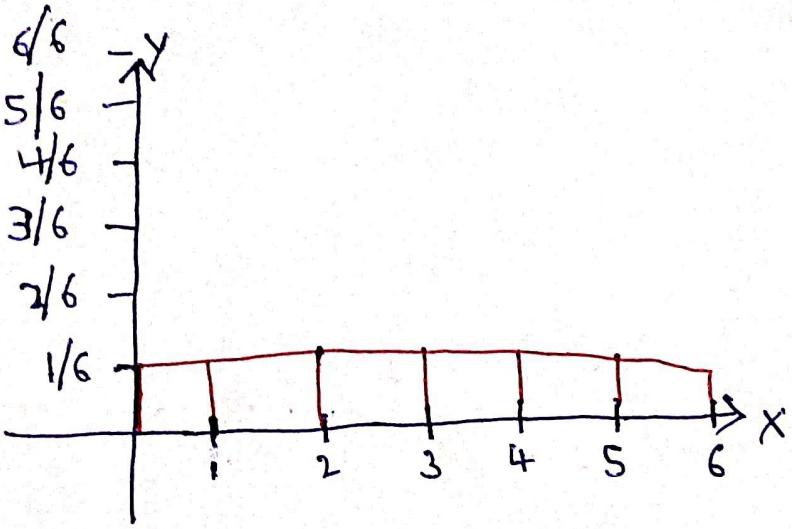
Probability Mass Function:

* used to Represent Discrete Random variables

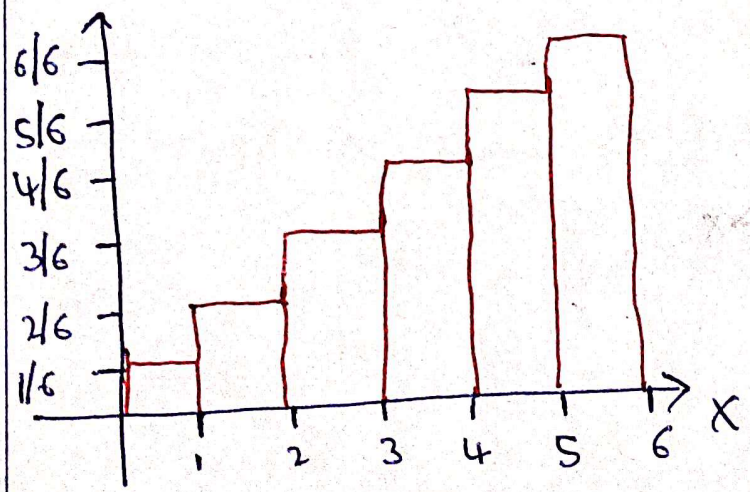
For eg: Rolling a Dice

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

$$Pr(1) = 1/6 \;, Pr(2) = 1/6 \;, \ldots$$



cumulative probability:

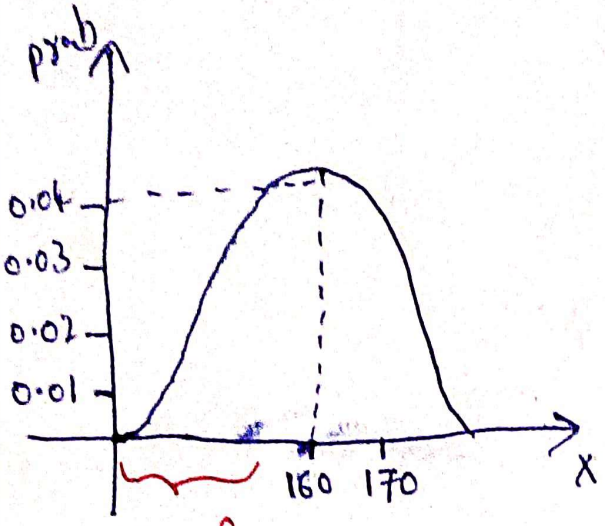


* It's the cumulative sum of all the Data points.

Probability Density Function:

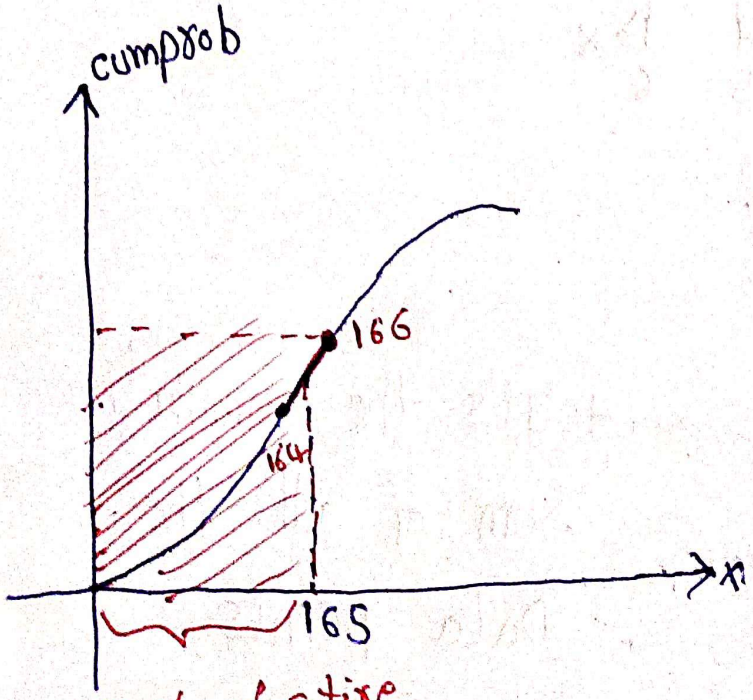
* used to represent continous Random variables.

Eg: Heights



50% of Entire Distribution

CDF:



50% of Entire

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

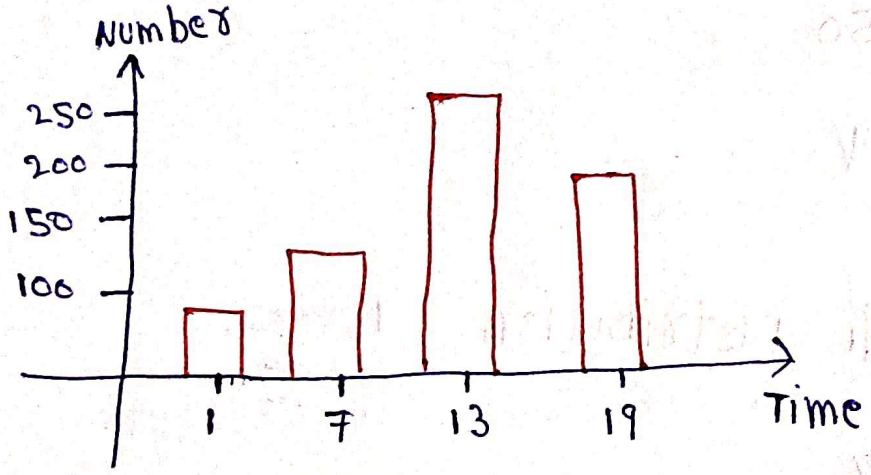
Variance : npq

std Deviation : $\sqrt{npq}$

Poisson Distribution:

* Describes the number of events occuring at a certain period of Time.

eg: No of people visiting Hospital every '6' hours.



$\lambda = 135$

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

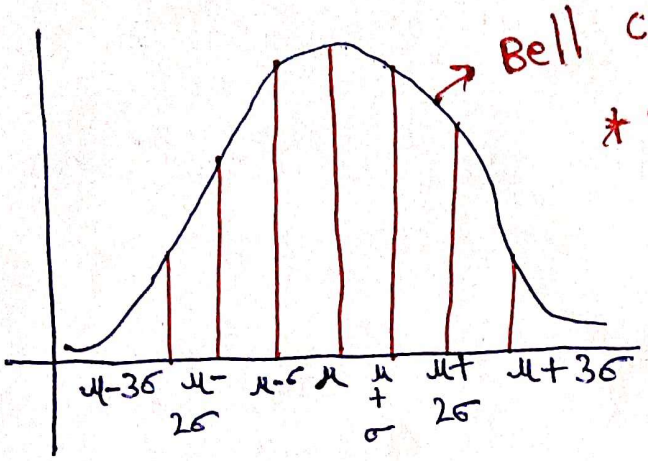
Mean $\rightarrow \lambda t$

Variance $\rightarrow \lambda t$

* t = time Interval

λ = Expected no. of Events

Normal (or) Gaussian Distribution:



* symetric Distribution

Empirical Rule :

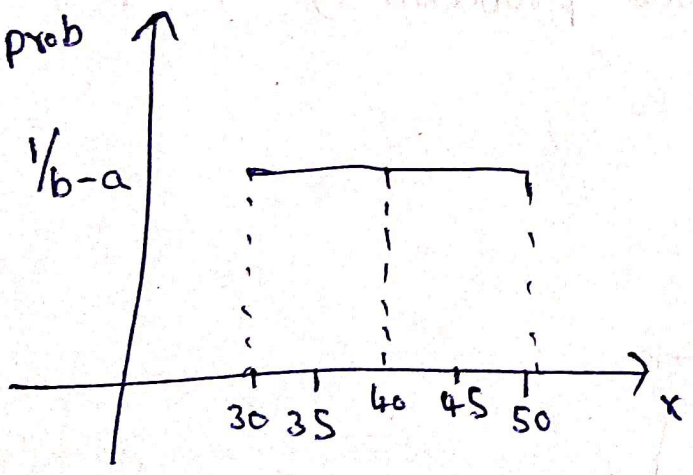
* 68% Data lies b/w $\mu-\sigma$ and $\mu+\sigma$
* 95% Data lies b/w $\mu-2\sigma$ and $\mu+2\sigma$
* 99.7% Data lies b/w $\mu-3\sigma$ and $\mu+3\sigma$

eg: Based on so much of Analysis, Researchers Found Height, weight, iris, etc... will mostly have guassian Distribution

Uniform Distribution

continous uniform Distribution:

* In uniform Distribution, the probability of getting outcomes is equal
* In continous uniform Distribution, In between a specified Range the probability of getting outcomes is same.



$$Pr = (x_2 - x_1) * \frac{1}{b-a}$$

a = lower bound

b = higher bount.

x_1, x_2 are datapoints.

$$Pr(X \text{ b/w } 40 \& 45) = (x_2 - x_1) * \frac{1}{b-a}$$

$$= (45 - 40) * \frac{1}{50-30}$$

$$= 5/20$$

$$= 1/4 \rightarrow 0.25 \rightarrow 25\%$$

$Pr(x \text{ b/w } 30 \text{ \& } 35) = (35-30) * \frac{1}{20}$

$= 0.25 \Rightarrow \boxed{25\%}$

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

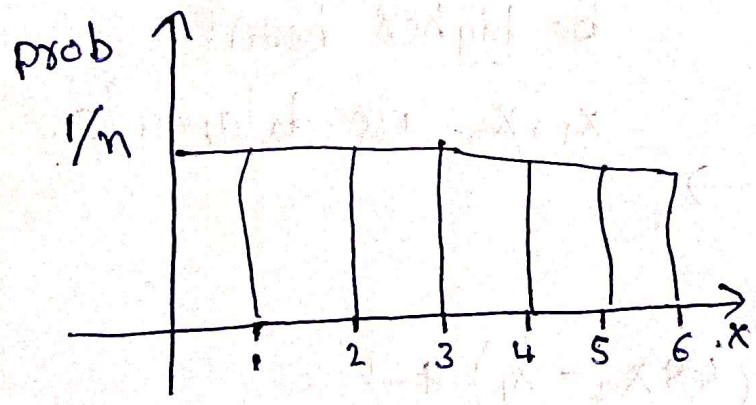
Mean : $a-b/2 = 15/20$

Variance : $(b-a)^2/12 = 0.75 \Rightarrow \boxed{75\%}$

Discrete uniform distribution:

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

eg: Rolling a die

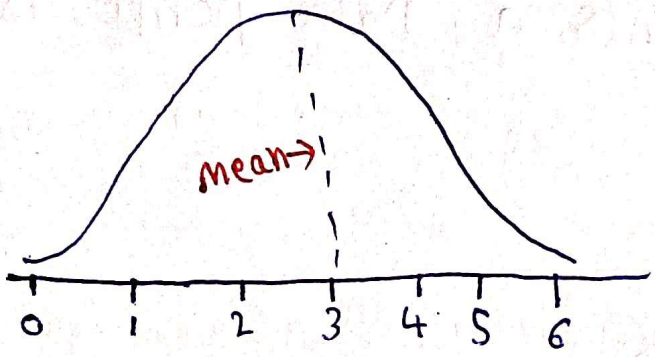


$Pr(1) = 1/n \Rightarrow 1/6$

standard Normal Distribution:

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

For eg this is Normal Distribution:



$\Rightarrow \mu = 0$

$\sigma = 1$

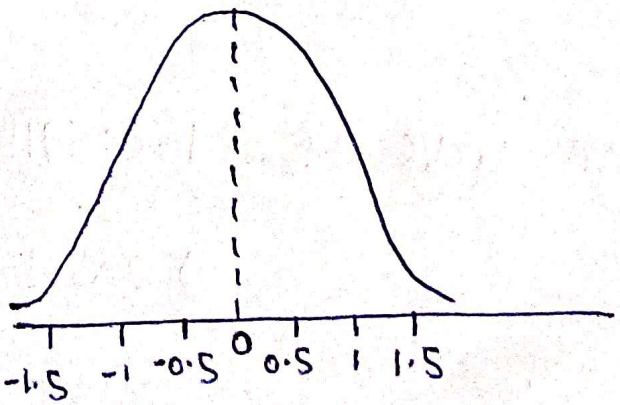
Let's Apply z score on all Data

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

μ = mean = 3

σ = std Deviation = 2

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



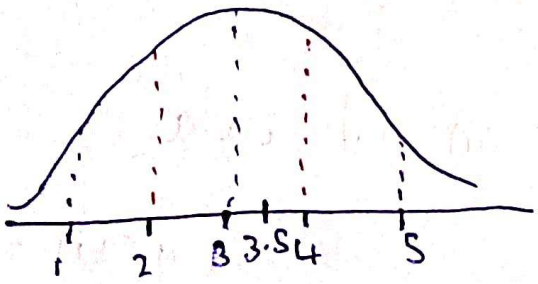
z-scores :

* 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.

eg :



* Find pr of Region above 3.5 ⇒

$z = \frac{3.5 - 3}{1}$ ⇒ 0.5 → Find it's value in z-Table

0.6915 } Its the region out of 1 Before 3.5

→ Region above 3.5 = 1 - 0.6915 ⇒ 0.3085

⇒ 30.85%

chi square:

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

$$x^2 = \frac{\Sigma(O-E)^2}{E}$$

O = observed data

E = Expected data

eg:

	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

$\alpha = 0.05$ (given)

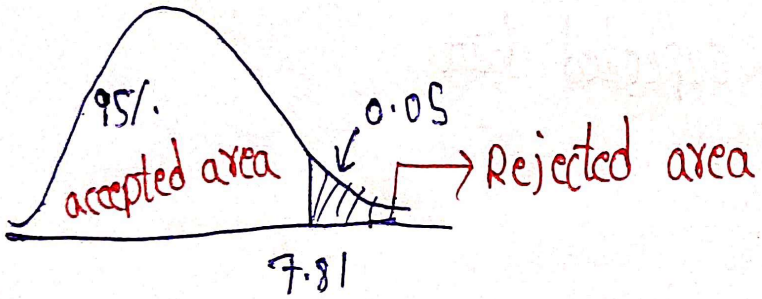
degree of freedom = n - 1

= 4 - 1

= 3

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 :

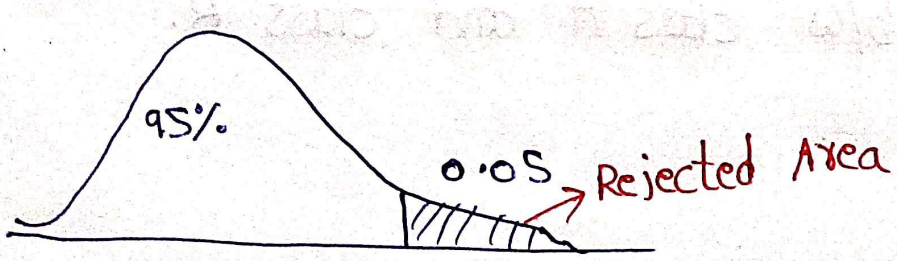
* It is used to find the comparision of variances of two data sets

$$F = \frac{\sigma(d1)^2}{\sigma(d2)^2}$$

eg :

class	size	variance
A	20	2.5
B	20	1.8

$\alpha = 0.05$



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

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

∴ As F value (1.39) < critical value (2.20) we can say there is no significant difference. b/w class A and class B.