Statistics

Introduction to Statistics: -

Stats Definition: - Stats is the science of collecting, organizing and analyzing data.

Data: - Facts or pieces of information

E.g.: - 1. Height of student in classroom

2. No. of sales in term of revenue of a company

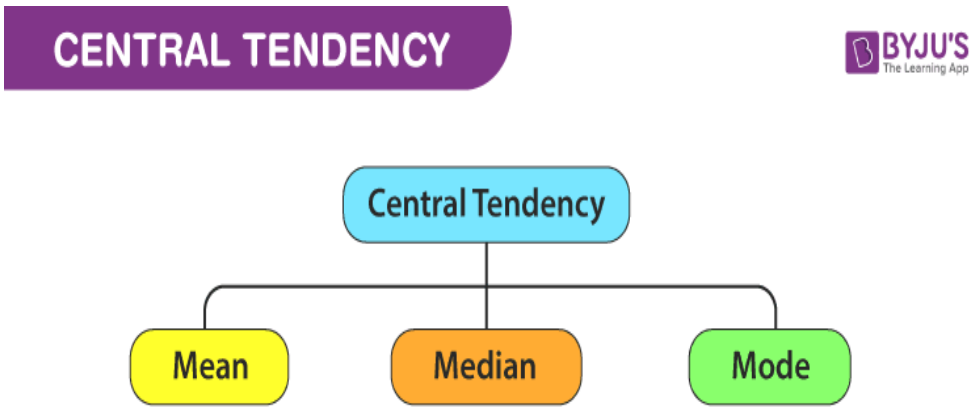
3. IQ of students in classroom

Type of Statistics: -

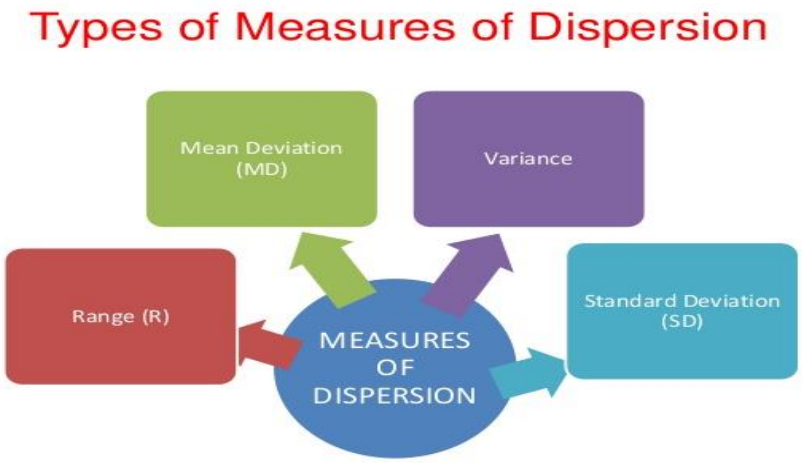
1. Descriptive Statistics
2. Inferential Statistics

1. **Descriptive Statistics: -** it consists of organizing summarizing and Visualizing data.

I. Measure of Central Tendency: -



II. Measures of Dispersion: -



III. Different type of distribution of data: -

i. Bernoulli Distribution
ii. Uniform Distribution
iii. Binomial Distribution
iv. Normal or Gaussian Distribution
v. Exponential Distribution
vi. Poisson Distribution

2. **Inferential Statistics**: - Inferential statistics are used to make conclusions about the population by using analytical tools on the sample data.

Measures of inferential statistics are

T-test
Z-test
CHI Square Test
Anova test
Hypothesis testing
P-Value
Significance value

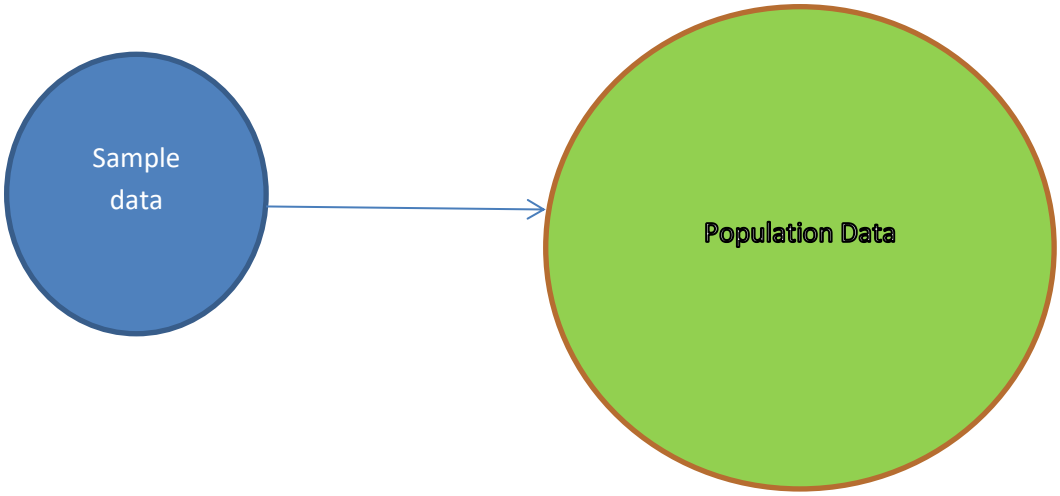
E.g.: - Let say there are 10 Cricket Camps in Bangalore and you have collected the height of cricketers from one of the camps.

Height is recorded are [175cm,180cm,140cm,140,135cm,160cm,135cm]

(Sample data)

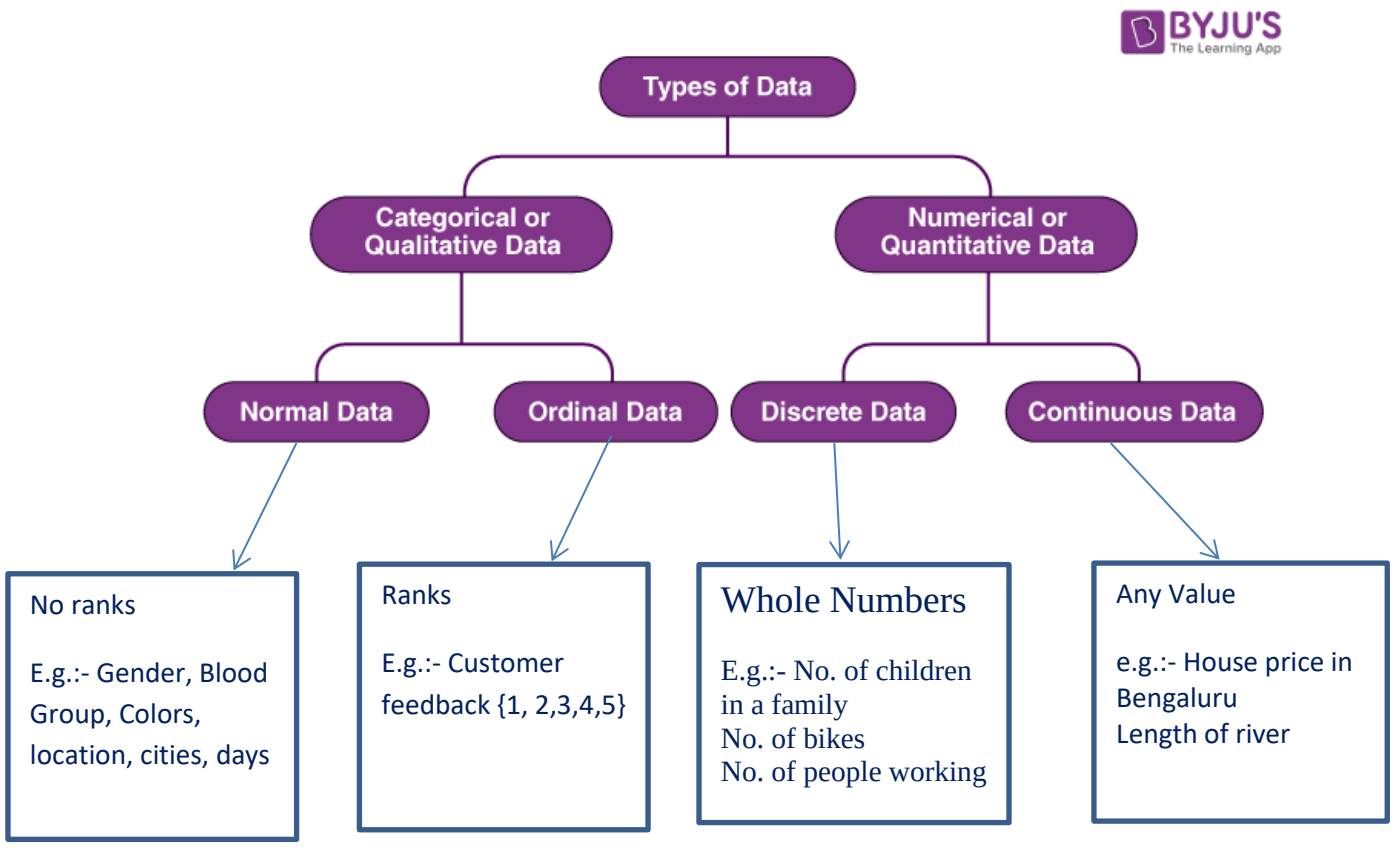
a. **Descriptive Question: -**
   IV. **What is the average height of the entire camps**
   V. **Disturbance of a data**
   VI. **140cm how many STD it is away from mean**

b. **Inferential Question: -**
   - **Are the average height of a players of camp1 similar to that of camp2**



➢ **Population and Sample data**: -
   - Population Data (N): - Population is a group or a superset of data that you are interested in studying.
   - Sample Data (n): - a sample is a subset of population data.

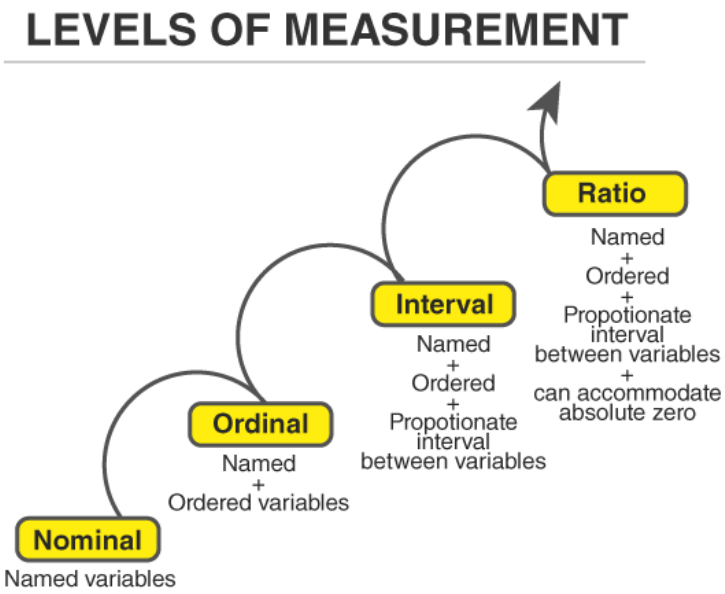
➢ **Types Of Data**: -



➢ **Scales of Measurement**: -the variables or numbers are defined and categorized using different **scales of measurements**. Each **level of measurement** scale has specific properties that determine the various use of statistical analysis

There are four different scales of measurement.

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale



I. Nominal Scale data: - A nominal scale is the 1st level of measurement scale in which the numbers serve as “tags” or “labels” to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value

- Qualitative/ Categorical Data
- E.g.: - Gender, color, Labels
- Order or rank does not matter

II. Ordinal Scale Data: - The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the “order.”

Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.

- Rank is important
- Order matters
- Difference cannot be measured
- **Example:**
    - Ranking of school students – 1st, 2nd, 3rd, etc.
    - Assessing the degree of agreement
        - Totally agree
        - Agree
        - Neutral
        - Disagree
        - Totally disagree

III. Interval Scale Data: - The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

- The order matters
- Difference can be measured
- The ratio cannot be measured
- No '0' starting point
- **Example:**
    - Likert Scale
    - Net Promoter Score (NPS)
    - Bipolar Matrix Table
    - IQ

IV. Ratio Scale Data: - The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio

scale has a unique feature. It possesses the character of the origin or zero points.

- The order matters
- Differences are measurable (Ratio)
- Contant a “0” Starting point
- E.g.: -
    - Students marks in a class

❖Descriptive Statistics

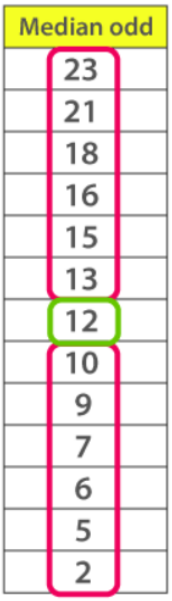
1. Measure of Central Tendency: -

- Mean
- Median
- Mode

➢ **Mean**: - The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values.

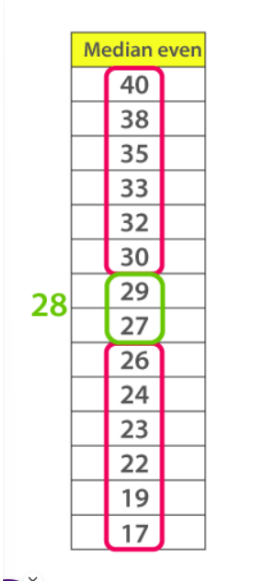
$$Mean = \frac{x_1+x_2+..+x_n}{n}$$

➢ **Median**: - Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values. Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2



Here 12 is the middle or median number that has 6 values above it and 6 values below it.

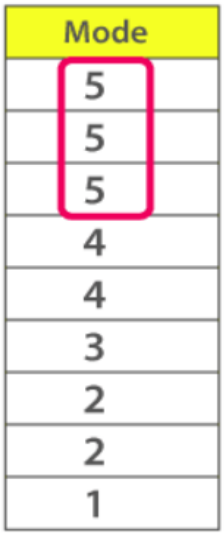
Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17



When you look at the given dataset, the two middle values obtained are 27 and 29. Now, find out the mean value for these two numbers.
i.e., (27+29)/2 =28
Therefore, the median for the given data distribution is 28.

- **Mode**: - The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and, in some cases, it does not contain any mode at all.

  Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5



Mode
5
5
5
4
4
3
2
2
1

  Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

2. Measures of Dispersion: -

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.

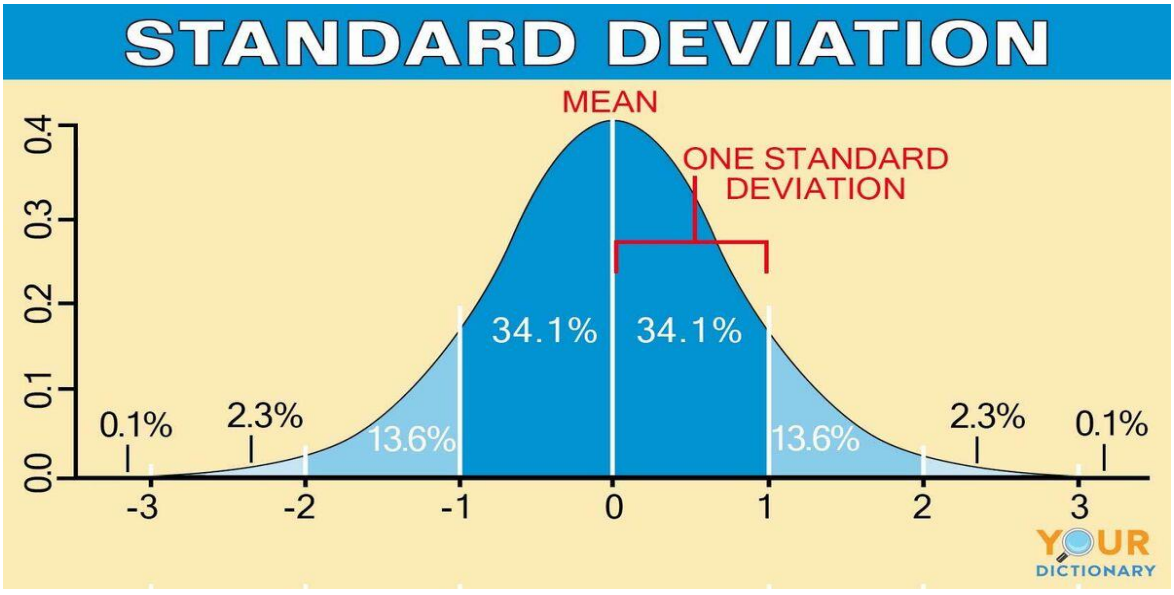
I. Variance: -

Population variance	Sample variance
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Here,	Here,
σ^2 = Variance	s^2 = Sample variance
x_i = ith observation of given data	x_i = ith observation of given data
μ = Population mean	$\bar{x}$ = Sample mean
N = Total number of observations (Population size)	n = Sample size (or Number of data values in sample)

- The sample variance is divided by n-1 so that we can create an Unbiased estimator of the population variance
- More the spread more the variance

II. Standard Deviation: - The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.

- A standard deviation is used to determine how estimations for a group of observations (i.e., data set) are spread out from the mean (average or expected value).
- How many STD Xi is away from mean



- **Random Variables: -** A random variable is a process of mapping the output of a random process or experiment to a number.

  E.g.: - Tossing a coin

  Rolling a dice

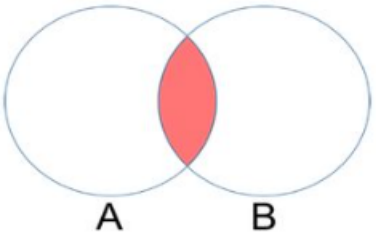
- **Sets: -**

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

  B= {3,4,5,6,7}

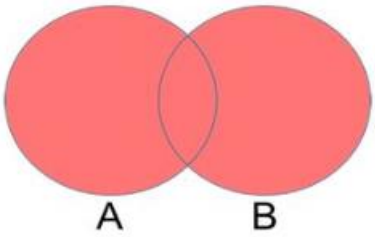
I. Intersection: -

$A \cap B = \{3,4,5,6,7\}$



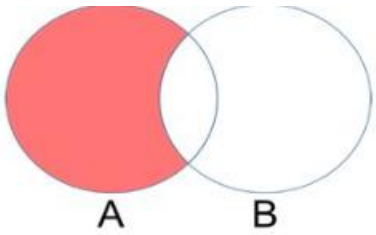
II. Union: -

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



III. Difference: -

A-B= {1,2,8}



IV. Subset: -

A → B = False

B → A= True

V. Superset: -

A ⟶ B = True

B ⟶ A= False

❖ Histograms and Skewness: -

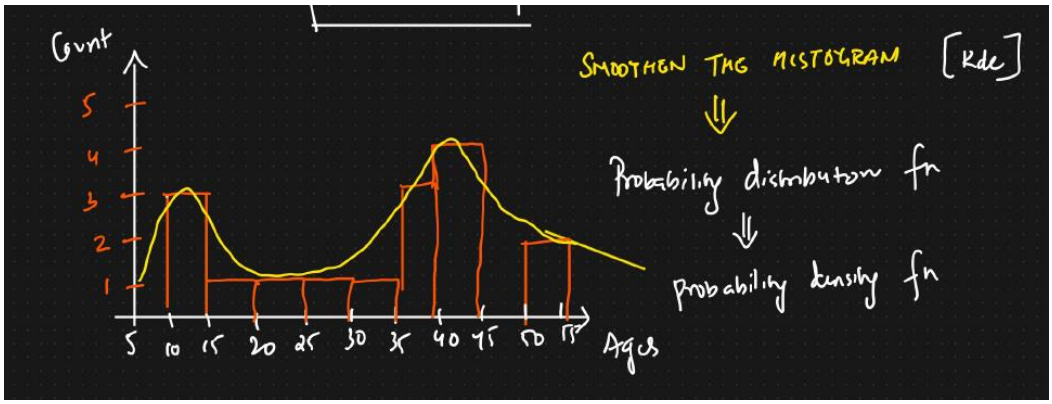
Histogram: -

Ages= {10,12,14,18,24,30,35,36,37,40,41,42,43,50,51}

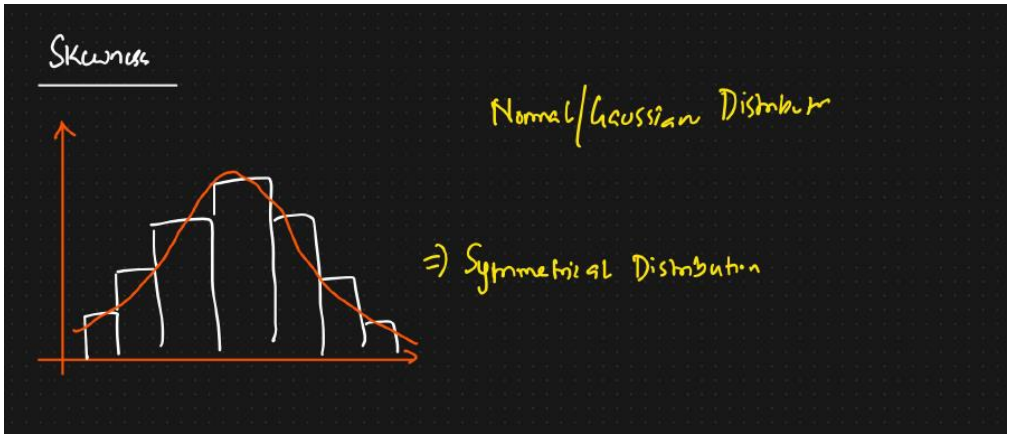
Bins, Bin size

No. of Bins=50/5=10

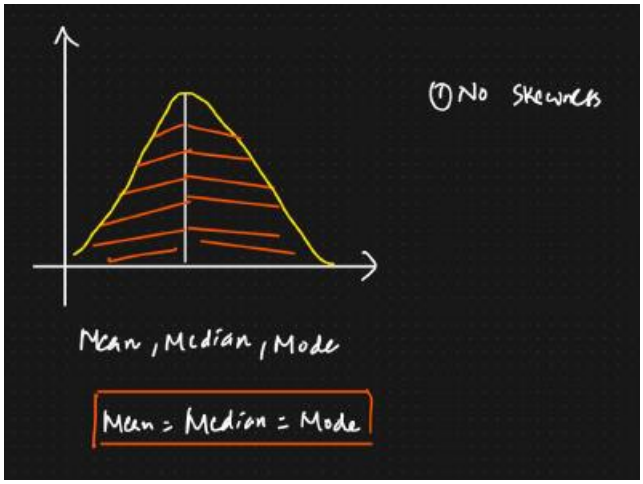
Bin size=5



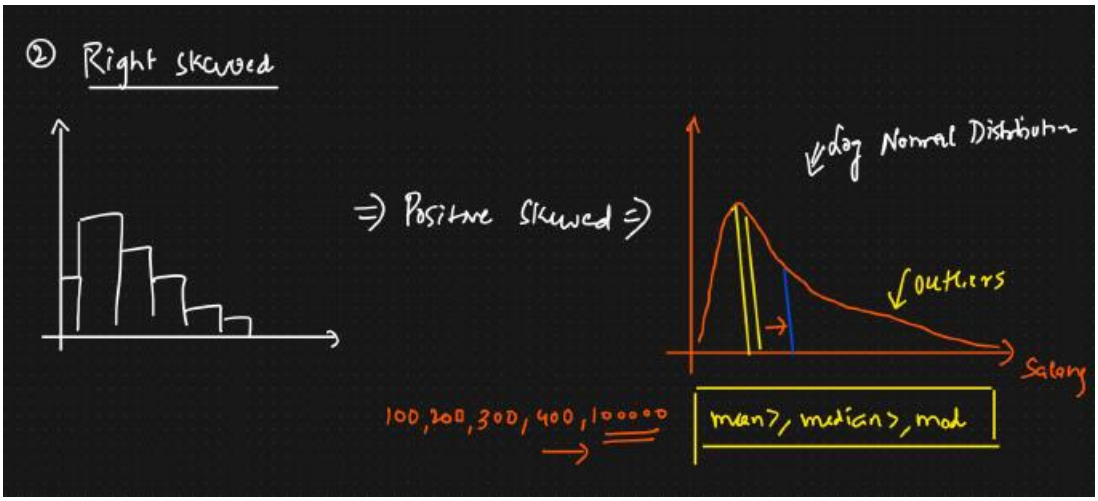
Skewness: - Skewness can be defined as a statistical measure that describes the lack of symmetry or asymmetry in the probability distribution of a dataset. It quantifies the degree to which the data deviates from a perfectly symmetrical distribution, such as a normal (bell-shaped) distribution. Skewness is a valuable statistical term because it provides insight into the shape and nature of a dataset's distribution.



A. No Skewed: -



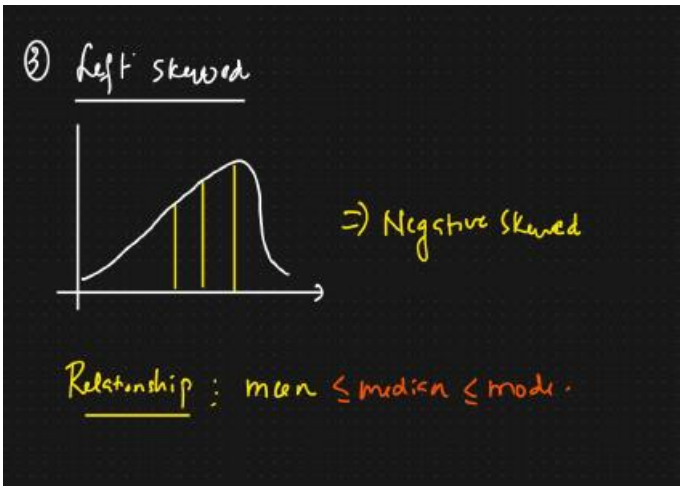
B. Right Skewed: -



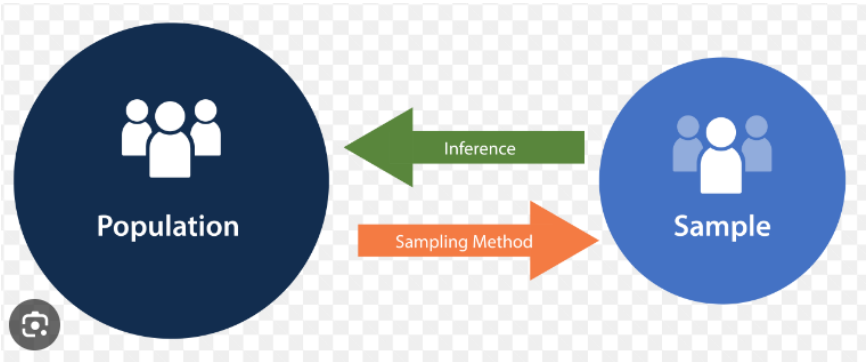
Mean > Median > Mode

C. Left Skewed: -

$$Mean < Median < Mode$$



❖ sampling Techniques: -



A. Simple random sampling:-

Example: Simple random sampling:- You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

B. Stratified sampling:-

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

C. Systematic sampling:-

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic sampling: - All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

D. Convenience sampling:-

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: Convenience sampling: - You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

E. Purposive sampling:-

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: Purposive sampling:- You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

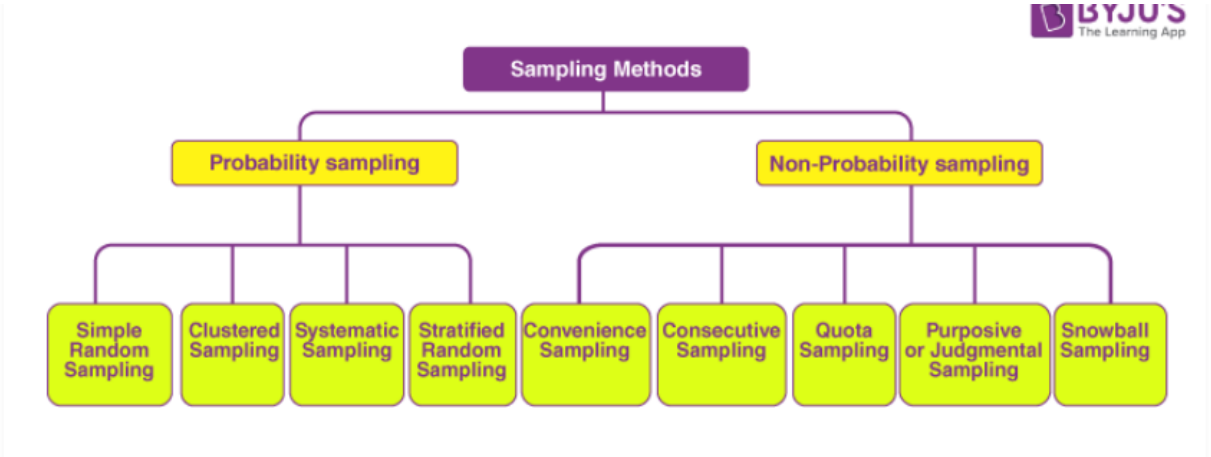
F. Cluster sampling:-

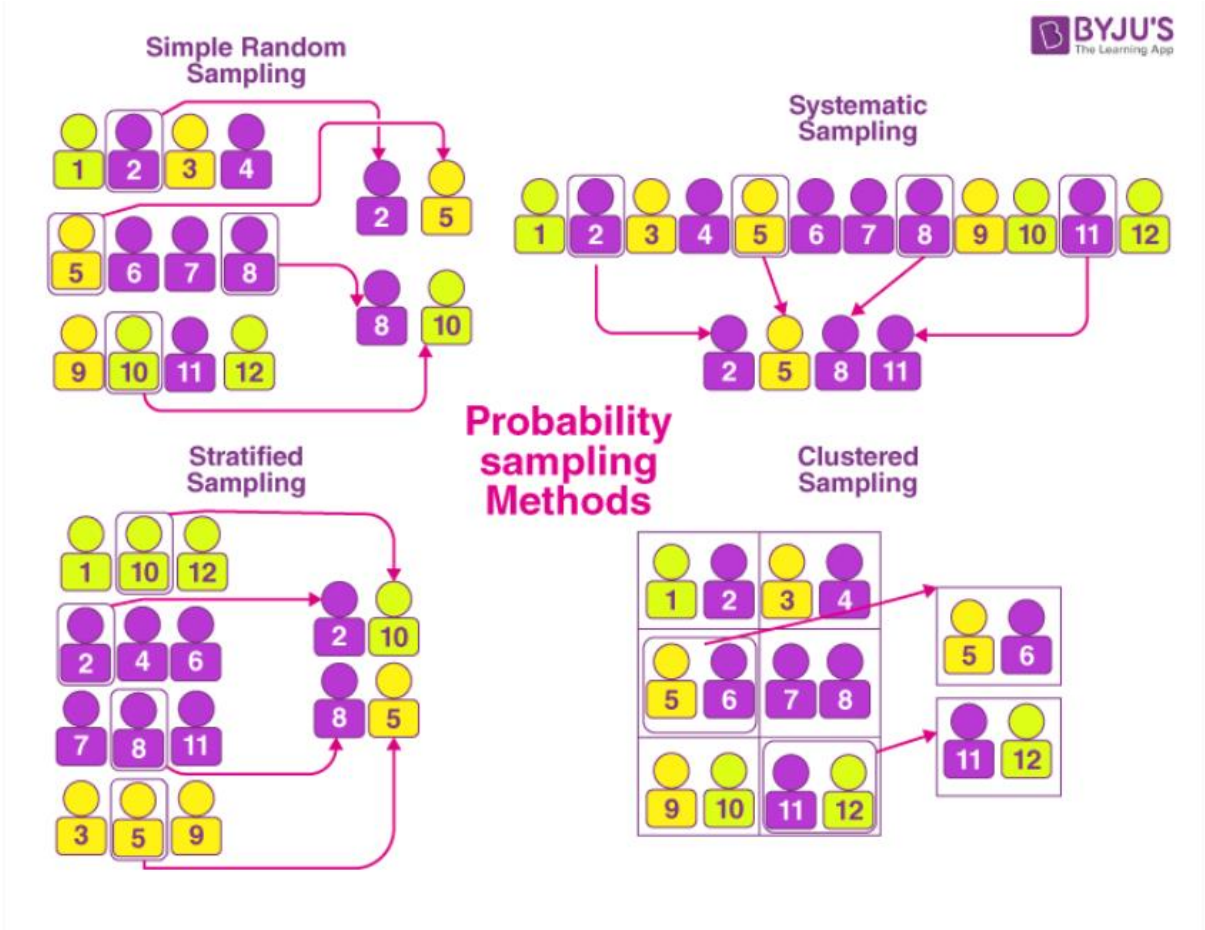
Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

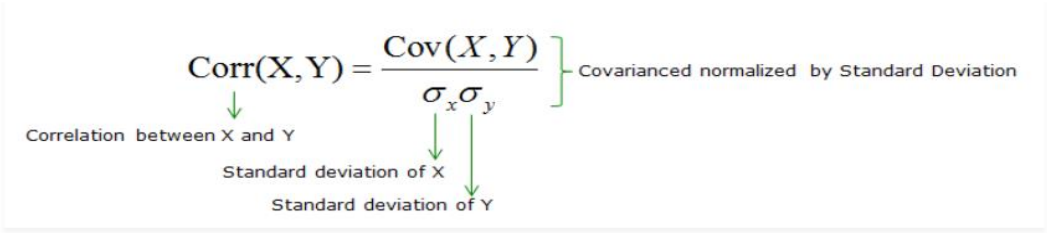
Example: Cluster sampling: - The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.


Sampling Methods
Probability sampling
Non-Probability sampling
Simple Random Sampling
Clustered Sampling
Systematic Sampling
Stratified Random Sampling
Convenience Sampling
Consecutive Sampling
Quota Sampling
Purposive or Judgmental Sampling
Snowball Sampling


Simple Random Sampling
Systematic Sampling
Probability sampling Methods
Stratified Sampling
Clustered Sampling

❖ Covariance and Correlation: -

- Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.
- The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.
- The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.
- A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.
- Positive: An increase in one of the variables results in an increase in the other.
- Negative: The variables are in opposite directions.
- Zero: Then, no relationship exists.



Covariance explains the joint variability of the variables.

$$\text{Cov(X,Y)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Where

x_i = Data value of x

y_i = Data value of y

$\bar{x}$ = Mean of x

$\bar{y}$ = Mean of y

N = Number of data values

A. Pearson correlation coefficient: - The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- **cov** is the covariance
- σx is the standard deviation of X
- σy is the standard deviation of Y

B. Spearman's rank correlation coefficient:- A correlation can easily be drawn as a scatter graph, but the most precise way to compare several **pairs of data** is to use a statistical test - this establishes whether the correlation is really significant or if it could have been the result of chance alone.
Spearman's Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables. The result will always be between 1 and minus 1.

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

❖ **Probability Distribution Function:** - a distribution function is a mathematical expression that describes the probability of different possible outcomes for an experiment.

Let us say we are running an experiment of tossing a fair coin. The possible events are **Heads**, **Tails**. And for instance, if we use X to denote the events, the probability distribution of X would take the value 0.5 for X=heads, and 0.5 for X=tails

- **Data Types**: - we have Qualitative and Quantitative data. And in Quantitative data, we have Continuous and Discrete data types.
  - **Continuous data** is measured and can take any number of values in a given finite or infinite range. It can be represented in decimal format. And the random variable that holds continuous values is called the Continuous random variable.

    Examples: A person's height, Time, distance, etc.

  - **Discrete data** is counted and can take only a limited number of values. It makes no sense when written in decimal format. And the random variable that holds discrete data is called the Discrete random variable.

    Example: The number of students in a class, number of workers in a company, etc.

- **Types of Probability Distributions**

  Two major kinds of distributions based on the type of likely values for the variables are,

  1. Discrete Distributions
  2. Continuous Distributions

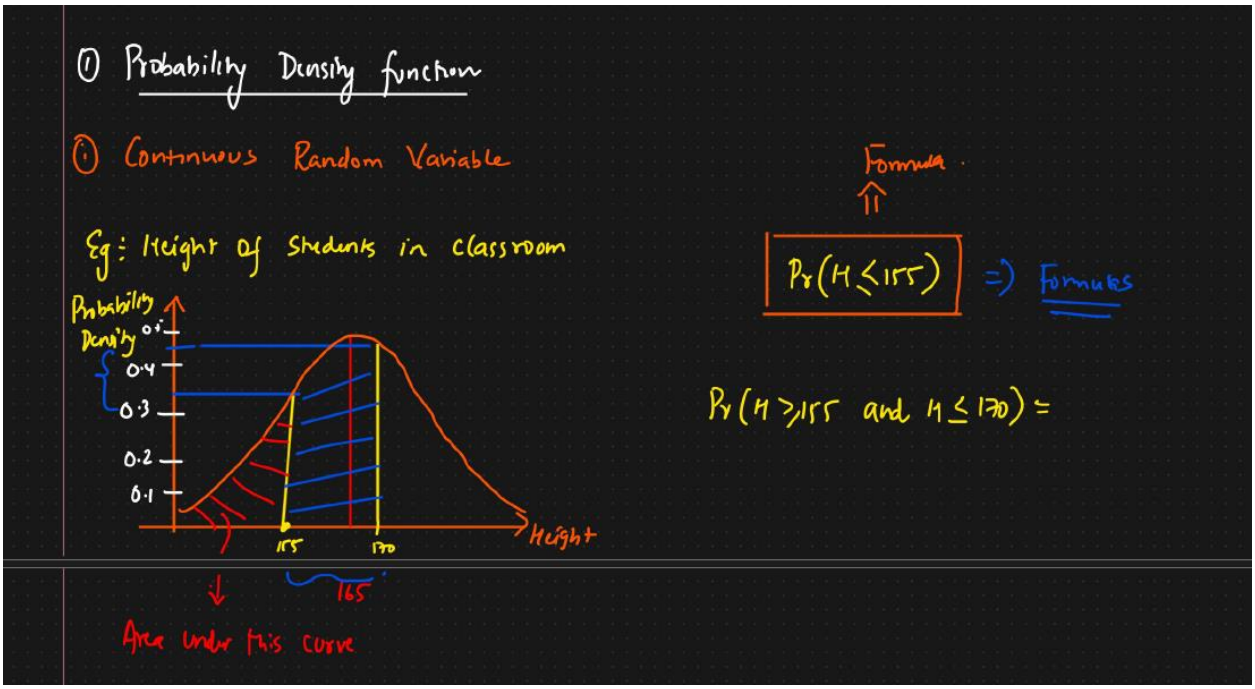
Discrete Distribution Vs Continuous Distribution

A comparison table showing difference between discrete distribution and continuous distribution is given here.

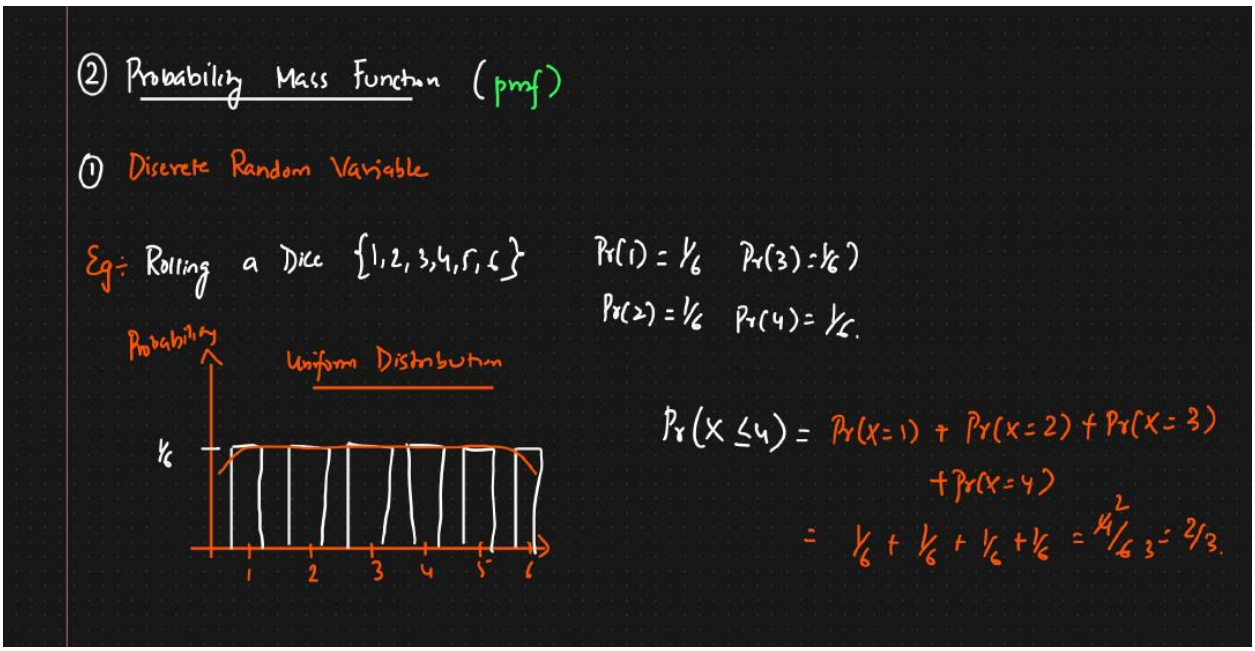
Discrete Distributions	**Continuous Distribution**
Discrete distributions have finite number of different possible outcomes	Continuous distributions have infinite many consecutive possible values
We can add up individual values to find out the probability of an interval	We cannot add up individual values to find out the probability of an interval because there are many of them
Discrete distributions can be expressed with a graph, piece-wise function or table	Continuous distributions can be expressed with a continuous function or graph
In discrete distributions, graph consists of bars lined up one after the other	In continuous distributions, graph consists of a smooth curve
Expected values might not be achievable	To calculate the chance of an interval, we required integrals

1. The probability distribution function / probability function has ambiguous definition. They may be referred to:
    - Probability density function (PDF)
    - Cumulative distribution function (CDF)
    - or probability mass function (PMF)
2. But what confirm is:
    - Discrete case: Probability Mass Function (PMF)
    - Continuous case: Probability Density Function (PDF)
    - Both cases: Cumulative distribution function (CDF)
3. Probability at certain x value, P(X=x) can be directly obtained in:
    - PMF for discrete case
    - PDF for continuous case
4. Probability for values less than x, P(X<x) or Probability for values within a range from a to b, P(a<X<b) can be directly obtained in:
    - CDF for both discrete / continuous case
5. Distribution function is referred to CDF or Cumulative Frequency Function

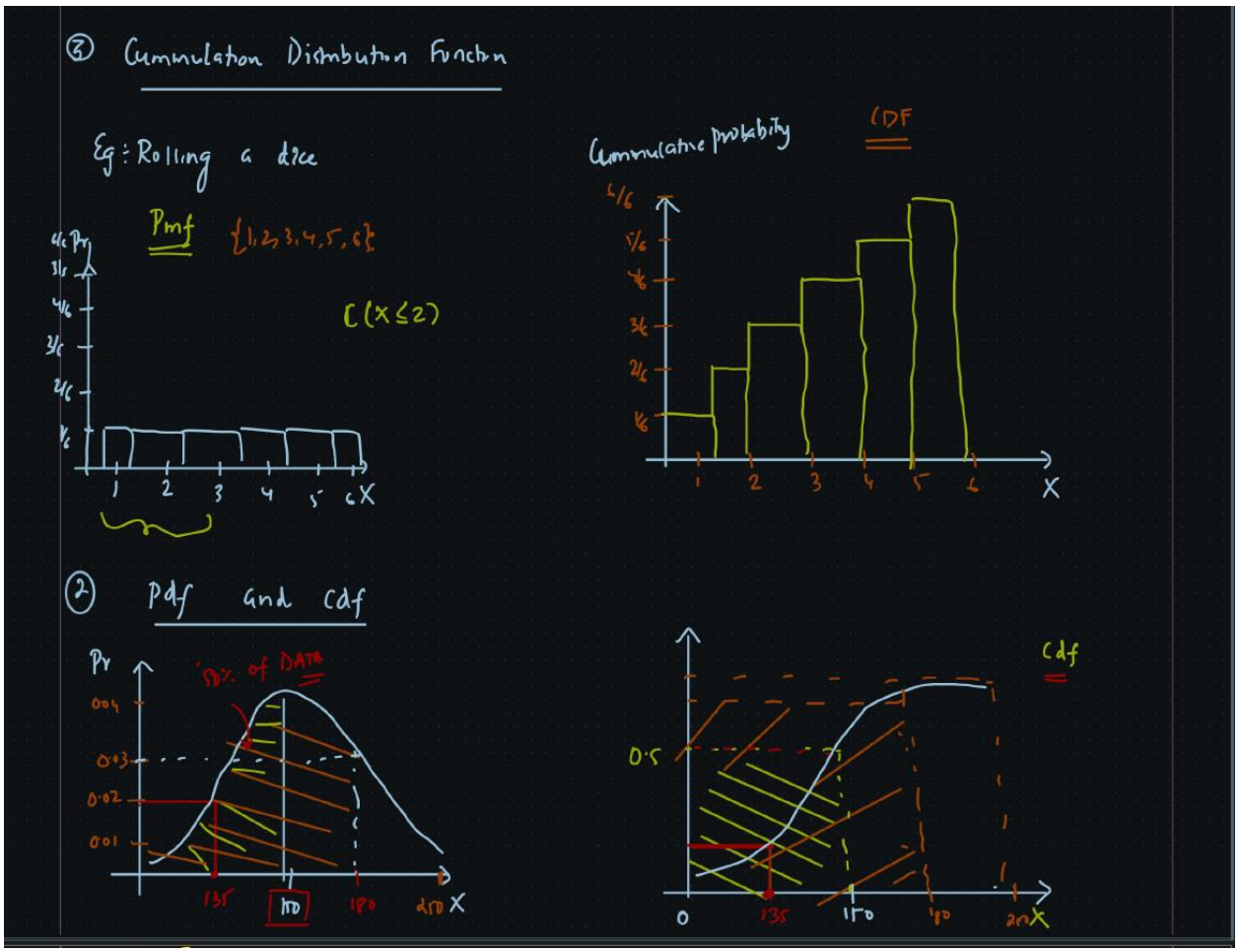
A. **Probability Density Function (PDF): -** It is a statistical term that describes the probability distribution of a continuous random variable. The probability associate with a single value is always Zero. Below is the formula for PDF.

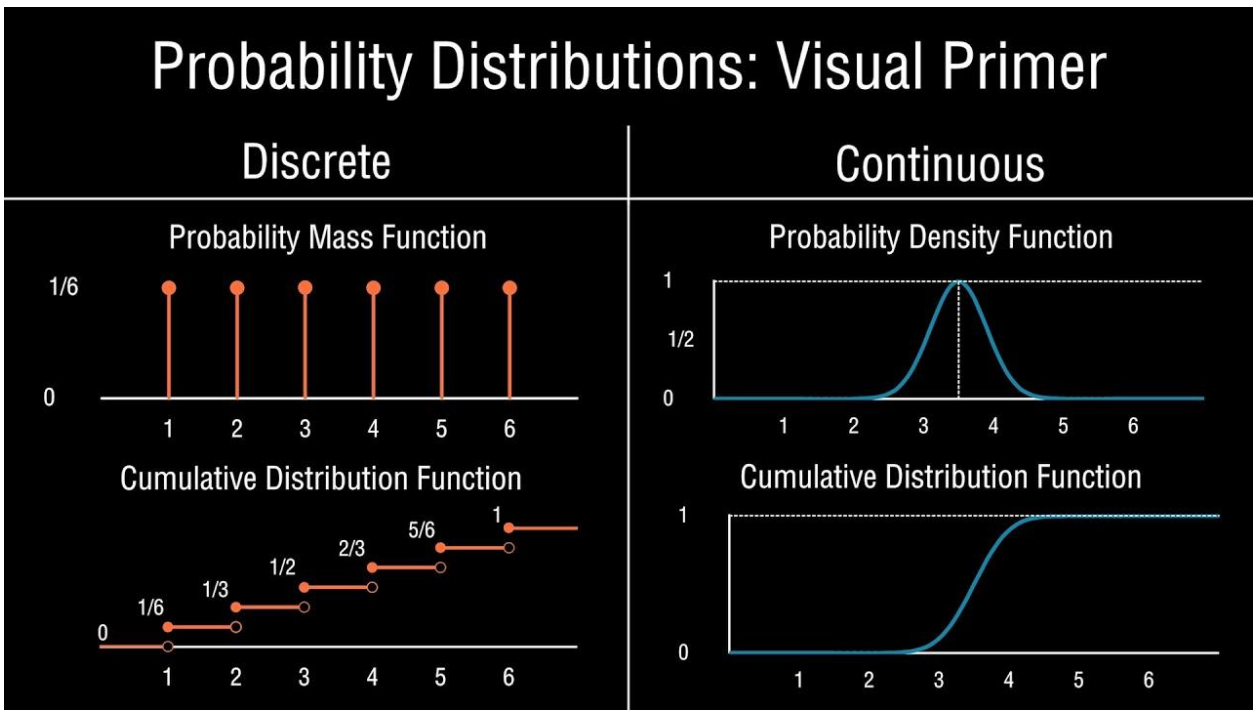


B. **Probability Mass Function (PMF):-** It is a statistical term that describes the probability distribution of a discrete random variable.



C. **Cumulative Distribution Function (CDF):-** It is another method to describe the distribution of a random variable (either continuous or discrete).





➢ Types of Probability Distribution: -

1. Normal or Gaussian Distribution
2. Bernoulli Distribution
3. Uniform Distribution
4. Poisson Distribution
5. Binomial Distribution
6. Log-Normal Distribution

1. Bernoulli Distribution: -

- Bernoulli distribution is a <u>discrete probability distribution</u>
- it's concerned with <u>discrete random variables {PMF}</u>
- Bernoulli distribution applies to events that have **one trial** and **two possible outcomes**. These are known as Bernoulli trials.

E.g.: -

- Tossing a coin {H,T}

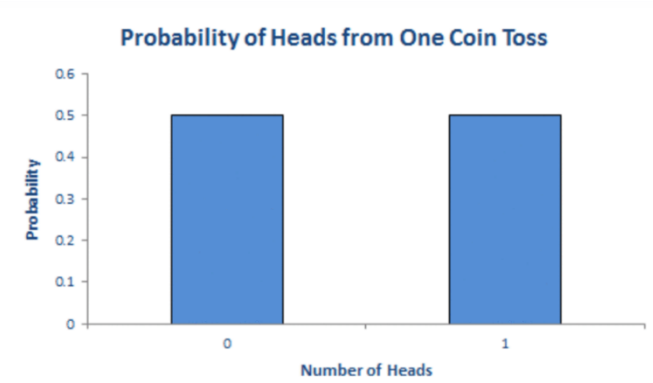
  P_r(H)=0.5 = p

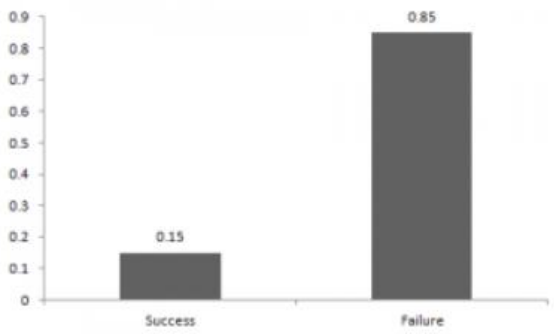
  P_r(T)=0.5 = 1-p=q

- Whether the person will Pass/Fail

  P_r(Pass)=0.85 = p

  P_r(Fail)= 1-p = 0.15 = q





PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0,1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1-p) = pq$

----→PMF=$P^k*(1-P)^{1-K}$

K{0,1} ----→ is outcomes

p→ Probability of one Outcome

q→ Probability of another Outcome

2. **Binomial Distribution:** -
   - it's concerned with <u>discrete random variables {PMF}</u>
   - There are two possible outcomes: true or false, success or failure, yes or no.
   - These Experiments is Performs for n trials
   - Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

E.g.: -

Tossing a Coin 10 times

$P(x:n,p) = {}^nC_x\, p^x\, (1-p)^{n-x}$

Or

$P(x:n,p) = {}^nC_x\, p^x\, (q)^{n-x}$

=PMF

$^{n}C_x = n!/x!(n-x)!$

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, ...

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

Mean, μ = np

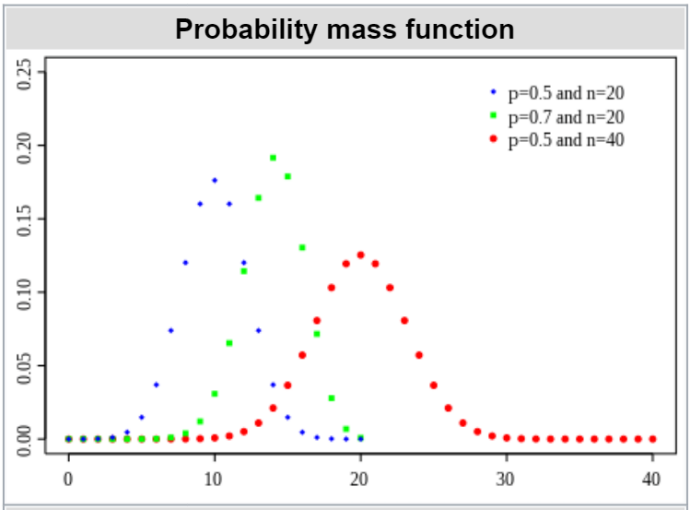
Variance, σ^2 = npq

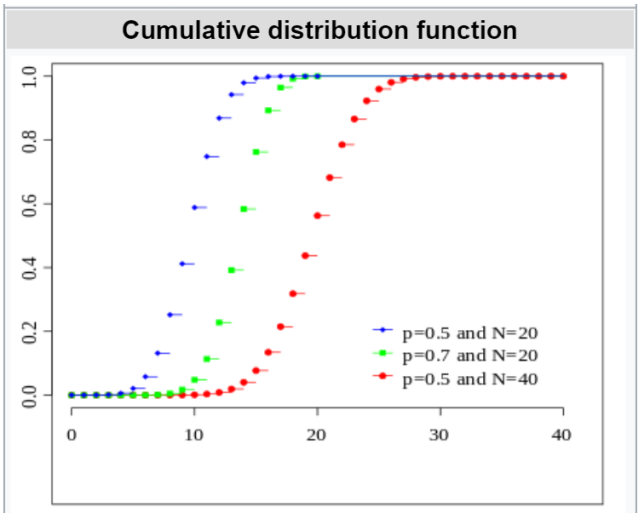
Standard Deviation σ= √(npq)

Where p is the probability of success

q is the probability of failure, where q = 1-p

Binomial distribution





3. Poisson Distribution: -

- it's concerned with <u>discrete random variables {PMF}</u>
- Describe the number of events occurring in a fixed time interval

E.g.: - No. of people visiting hospital every hour
No. of people visiting bank at 11am

$P(x, \lambda) = (e^{-\lambda} \lambda^x)/x!$

Where,

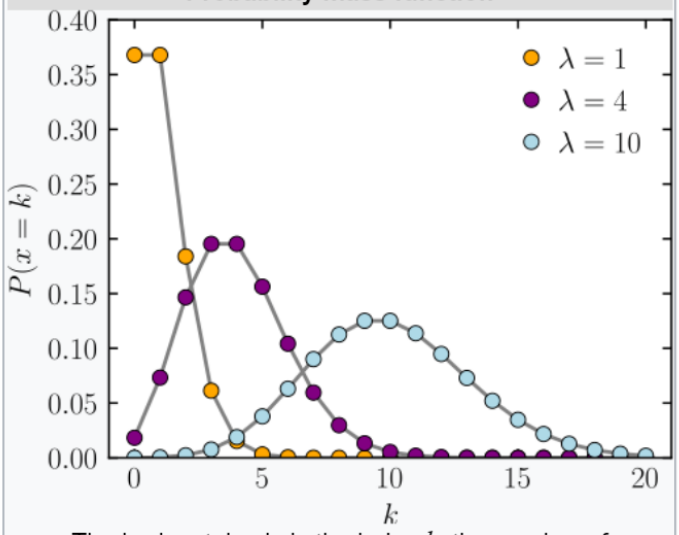
e is the base of the logarithm

x is the number of occurrences (x=0,1,2,…..)

λ Expected no. of events occur at

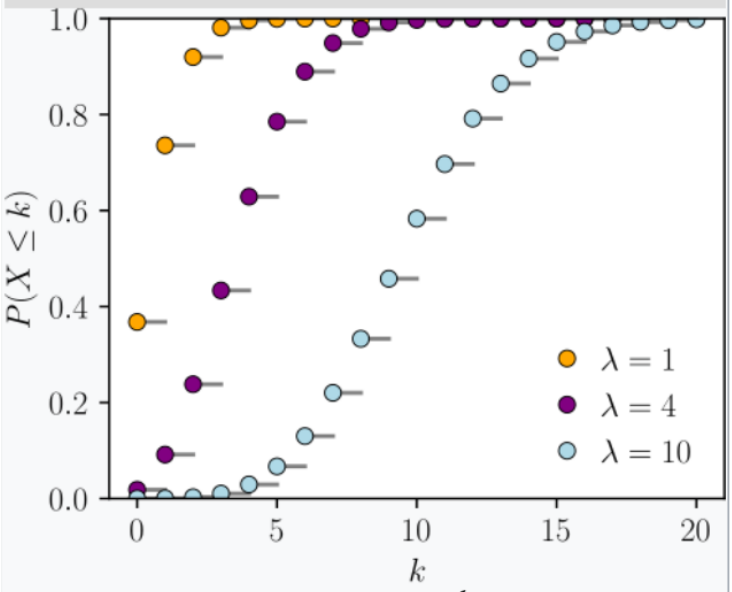
Poisson Distribution

Probability mass function



The horizontal axis is the index k, the number of occurrences. λ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given λ. The function is defined only at integer values of k; the connecting lines ar

Cumulative distribution function



The horizontal axis is the index k, the number of occurrences. The CDF is discontinuous at the integers of k and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

Notation	$\text{Pois}(\lambda)$
Parameters	$\lambda \in (0, \infty)$ (rate)
Support	$k \in \mathbb{N}$ (Natural numbers starting from 0)
PMF	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{j=0}^{\lfloor k \rfloor} \frac{\lambda^j}{j!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \left\lfloor \lambda + \frac{1}{3} - \frac{1}{50\lambda} \right\rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ

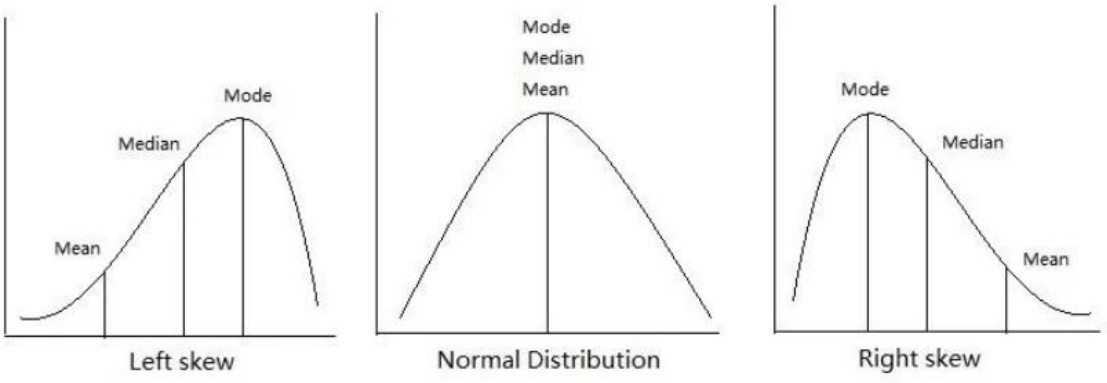
every time interval

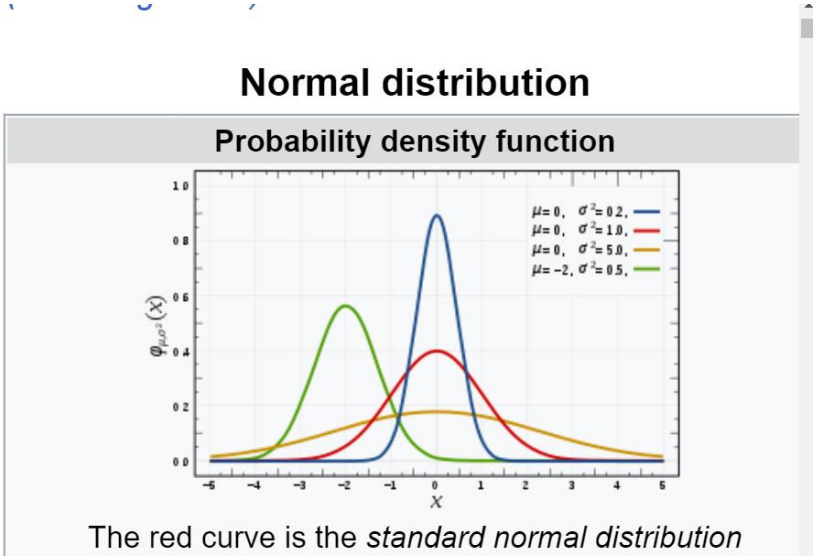
4. Normal or Gaussian Distribution: -

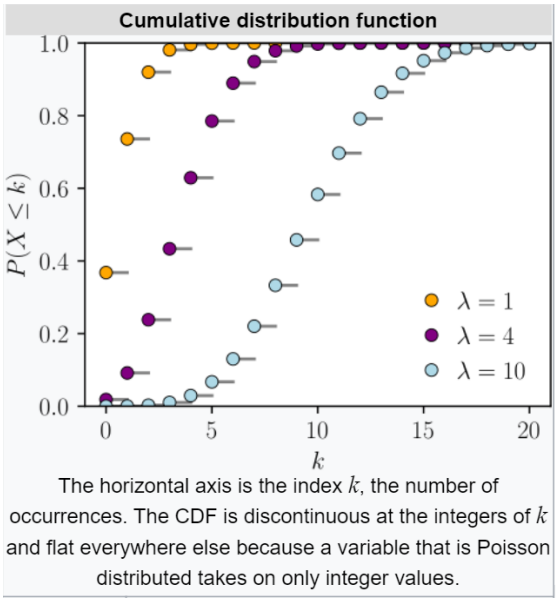
- it's concerned with Continuous random variables {PDF}
- Normal distributions are symmetrical, but not all symmetrical distributions are normal

Characteristics of Normal Distribution

- mean = median = mode
- Symmetrical about the center
- Unimodal
- 50% of values less than the mean and 50% greater than the mean







Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\,\operatorname{erf}^{-1}(2p-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2

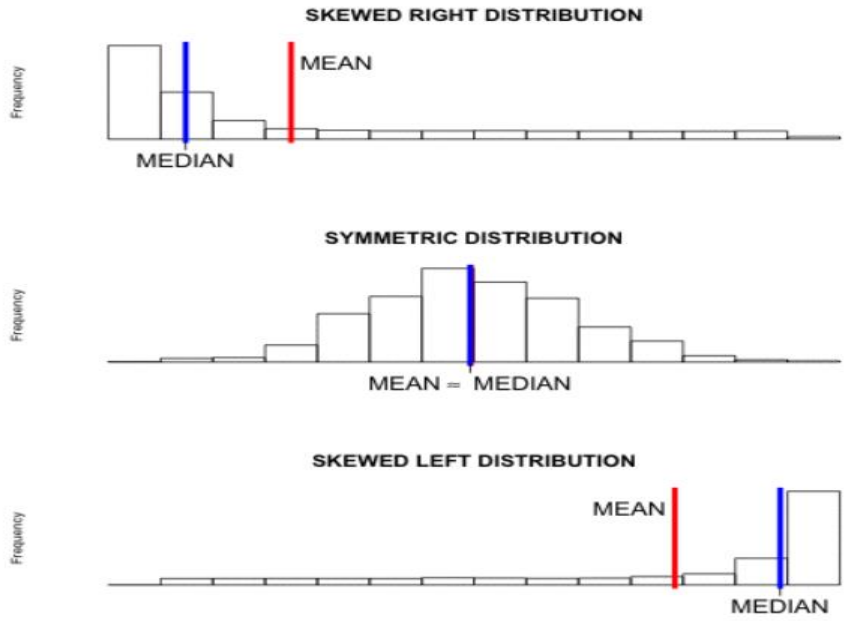
Here, *x* is value of the variable; *f(x)* represents the probability density function; μ *(mu)* is the mean; and σ *(sigma)* is the standard deviation.

Examples that mainly follow a Normal Distribution

1. Blood pressure
2. Height of students in a class
3. Errors while taking measurements
4. Marks in a test, etc

Some Basic Terminology

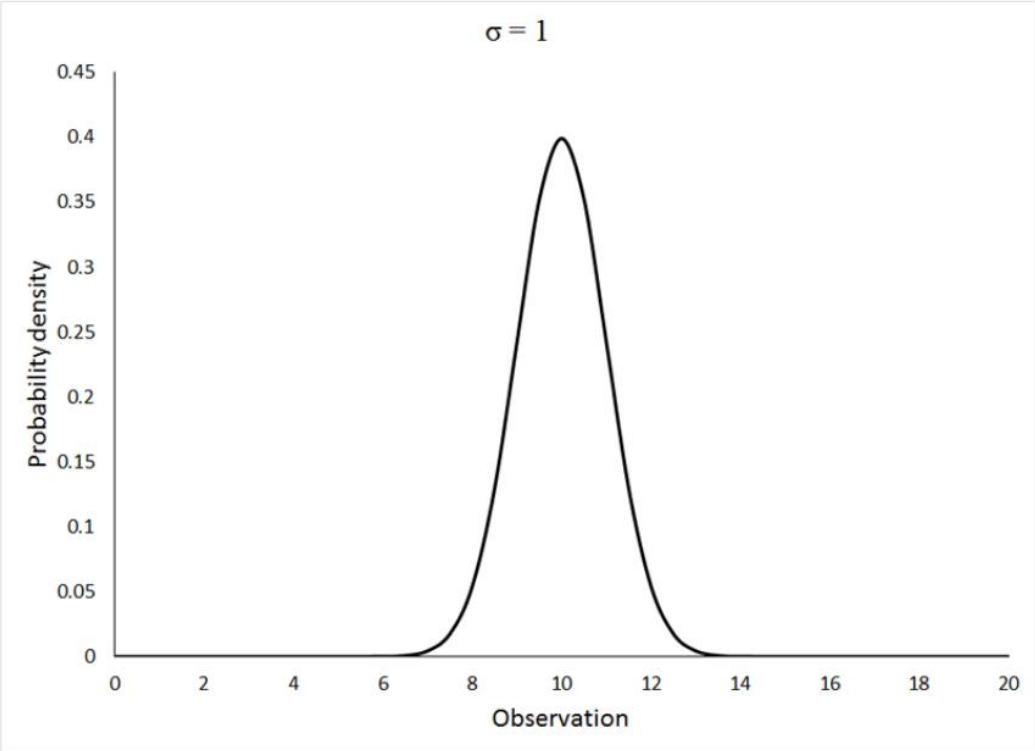
1. **Mean(μ)** — is the average of a data set.
2. **Median** — is the middle of the set of numbers.
3. **Mode** — is the most common number(peak) in a data set. A unimodal distribution only has one peak in the distribution, a bimodal distribution has two peaks, and a multimodal distribution has three or more peaks.

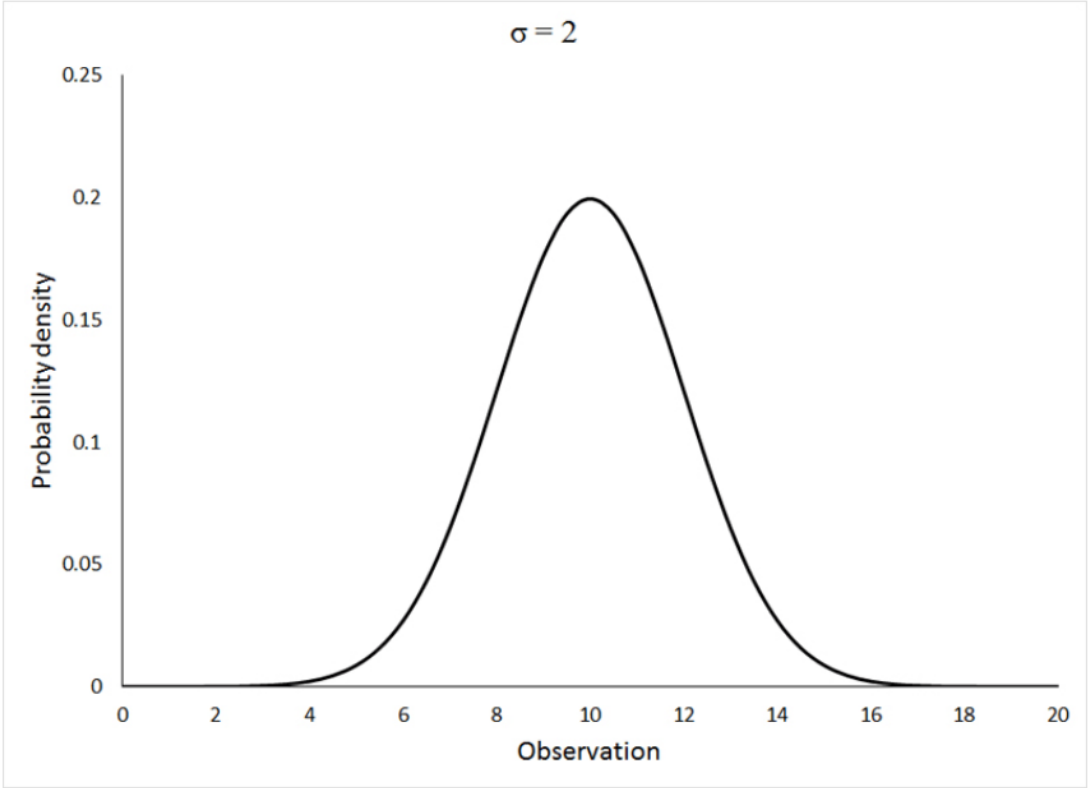


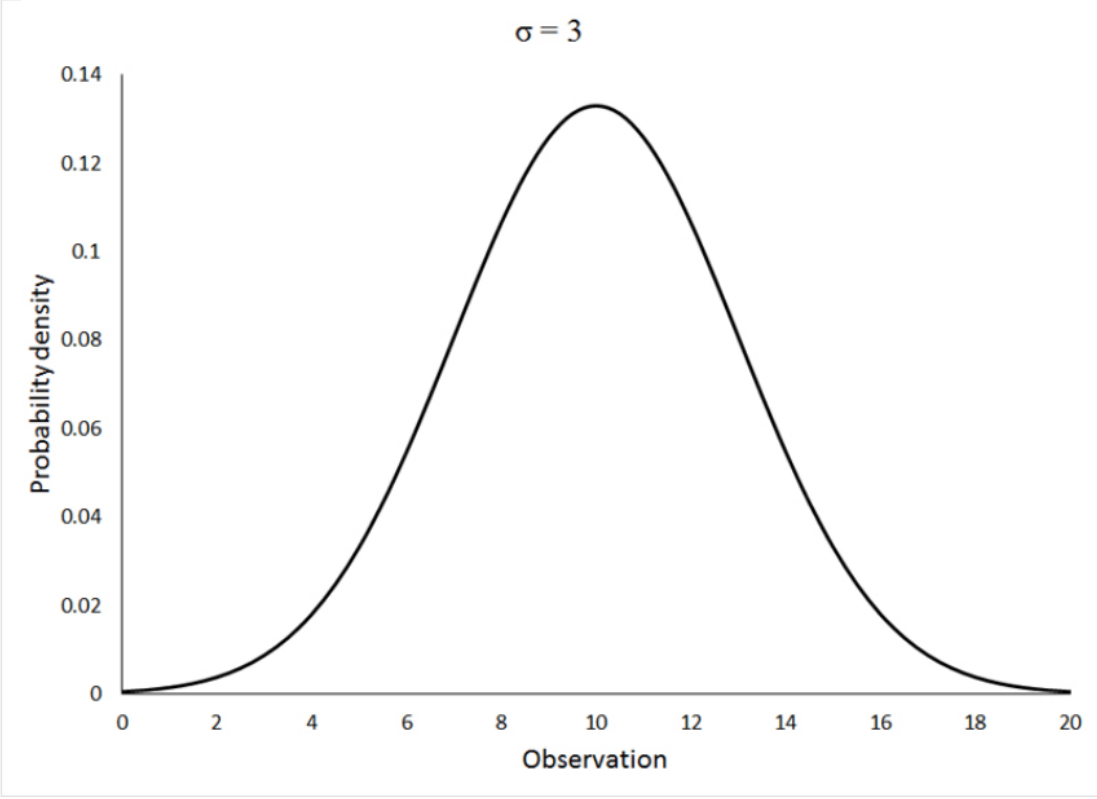
4. **Bias** — is the tendency of a statistic to overestimate or underestimate a parameter.

5. **Skewness** — refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.

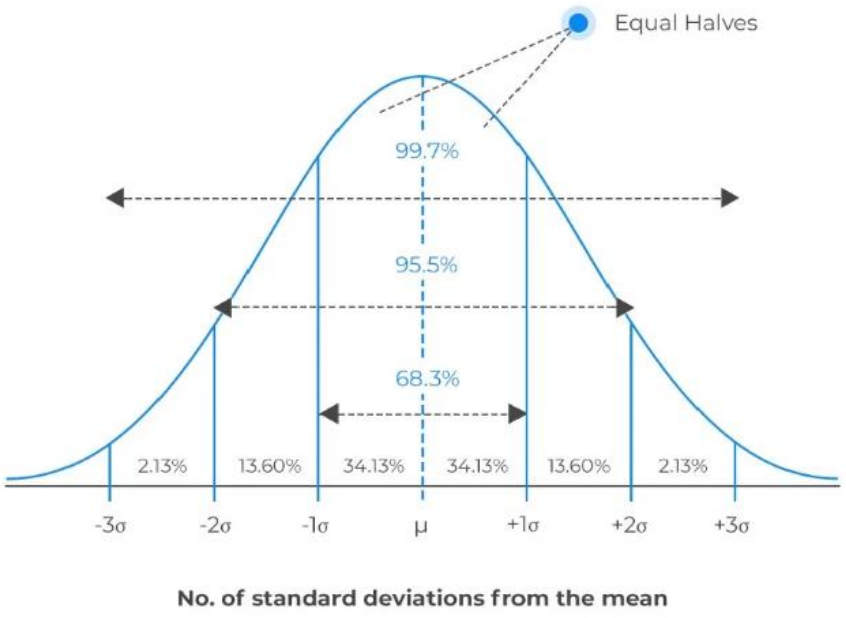
6. **Standard deviation(σ)** — is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.




σ = 2
0.25
0.2
0.15
0.1
0.05
0
Probability density
0
2
4
6
8
10
12
14
16
18
20
Observation


σ = 3
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability density
0
2
4
6
8
10
12
14
16
18
20
Observation

- **Empirical Rule of Normal Distribution: -** The empirical rule in statistics, also known as the 68 95 99 rule, states that for normal distributions, 68% of observed data points will lie inside one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will occur within three standard deviations.



- **68.3%** of values are within **1 standard deviation** (1σ) of the mean
- **95.5%** of values are within **2 standard deviations** (2σ) of the mean
- **99.7%** of values are within **3 standard deviations** (3σ) of the mean

It is always good to know the standard deviation because we can say that any value is:

- **likely** to be within 1 standard deviation (1σ)(68.3 out of 100 should be)
- **very likely** to be within 2 standard deviations (2σ) (95.5 out of 100 should be)

- **almost certainly** within 3 standard deviations (3σ) (997 out of 1000 should be)

5. Uniform Distribution: -

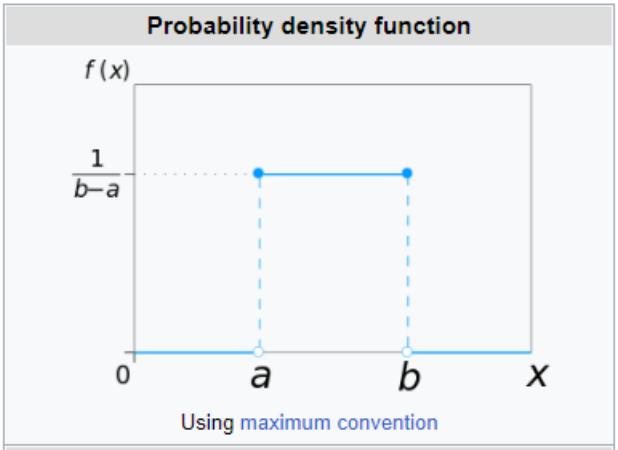
I. Continuous Uniform Distribution (PDF)

II. Discrete Uniform Distribution (PMF)

I. Continuous Uniform Distribution (PDF): -

- Continuous random variables {PDF}

Continuous uniform distribution with parameters a and b



Notation	$\mathcal{U}_{[a,b]}$
Parameters	$-\infty < a < b < \infty$
Support	$[a, b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$
Mean	$\frac{1}{2}(a+b)$
Median	$\frac{1}{2}(a+b)$
Mode	any value in (a, b)
Variance	$\frac{1}{12}(b-a)^2$

Eg: The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 and a minimum of 10.

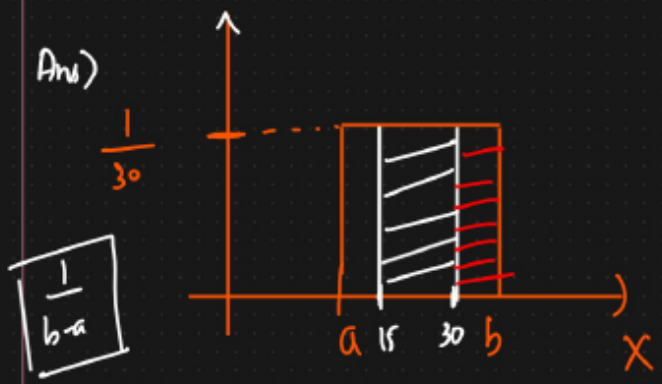
(i) Probability of daily sales to fall between 15 and 30?

$x_1 = 15$

$x_2 = 30$

$a = 10$ $\quad b = 40$

Ans)



$$Pr(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$$

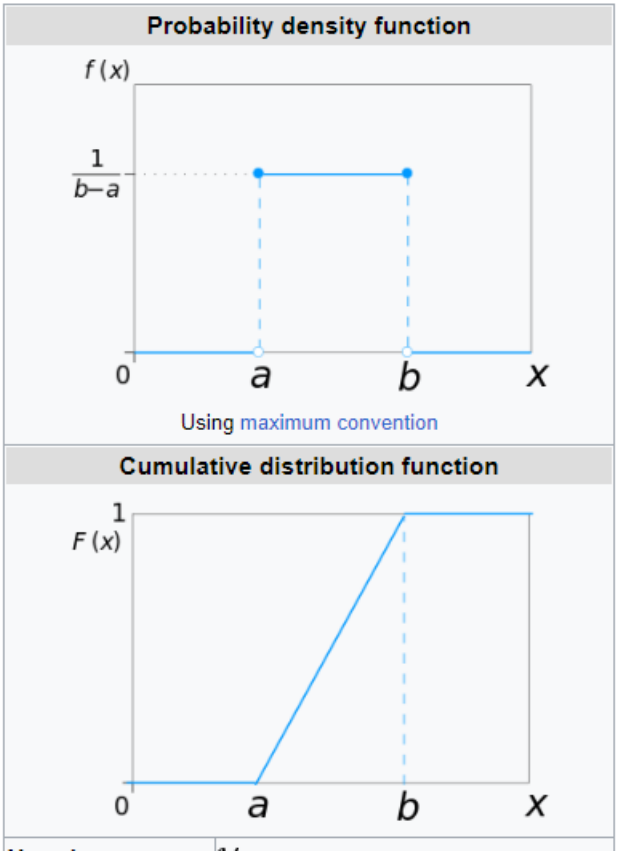
$$= 15 * \frac{1}{30} = \frac{1}{2} = 0.5$$

$$Pr(x \ge 30) = (40-30) \times \frac{1}{30} = 10 \times \frac{1}{30} = \frac{1}{3} = 0.33 = 33\%$$

II. Discrete Uniform Distribution (PMF): -

- Discrete random variables {PMF}

Continuous uniform distribution with parameters a and b



Notation	$\mathcal{U}\{a, b\}$ or $\operatorname{unif}\{a, b\}$
Parameters	a, b integers with $b \geq a$ $n = b - a + 1$
Support	$k \in \{a, a+1, \ldots, b-1, b\}$
PMF	$\frac{1}{n}$
CDF	$\frac{\lfloor k \rfloor - a + 1}{n}$
Mean	$\frac{a+b}{2}$
Median	$\frac{a+b}{2}$
Mode	N/A
Variance	$\frac{n^2-1}{12}$

➢ **Standard Normal Distribution Z-Score: - The standard normal distribution is a specific type of normal distribution where the mean is equal to 0 and the standard deviation is equal to 1.**

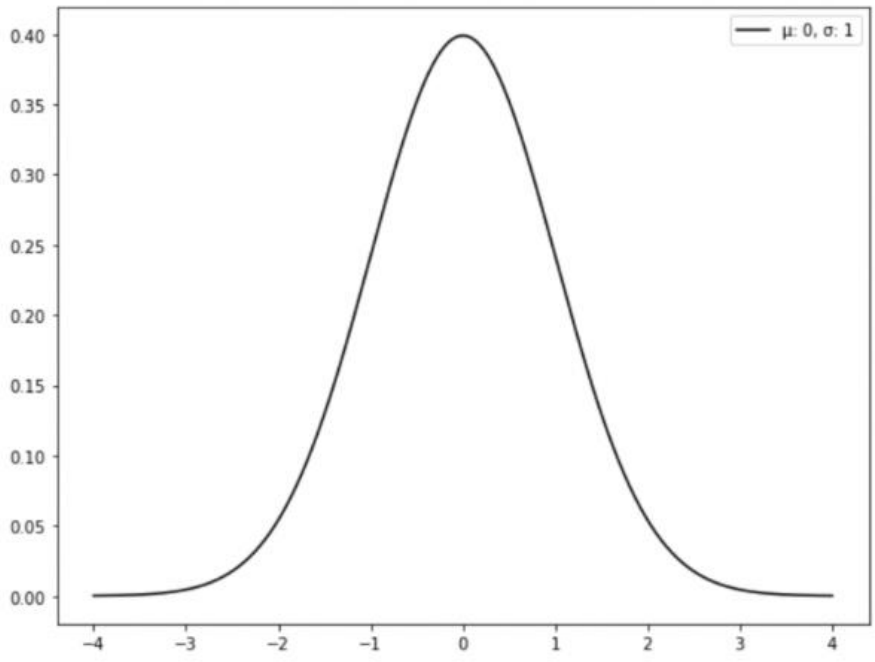
The normal distribution is the most commonly used probability distribution in statistics.

It has the following properties:

- Symmetrical
- Bell-shaped
- Mean and median are equal; both located at the center of the distribution

The mean of the normal distribution determines its location and the standard deviation determines its spread.

The following plot shows a standard normal distribution:



A standard normal distribution has the following properties:

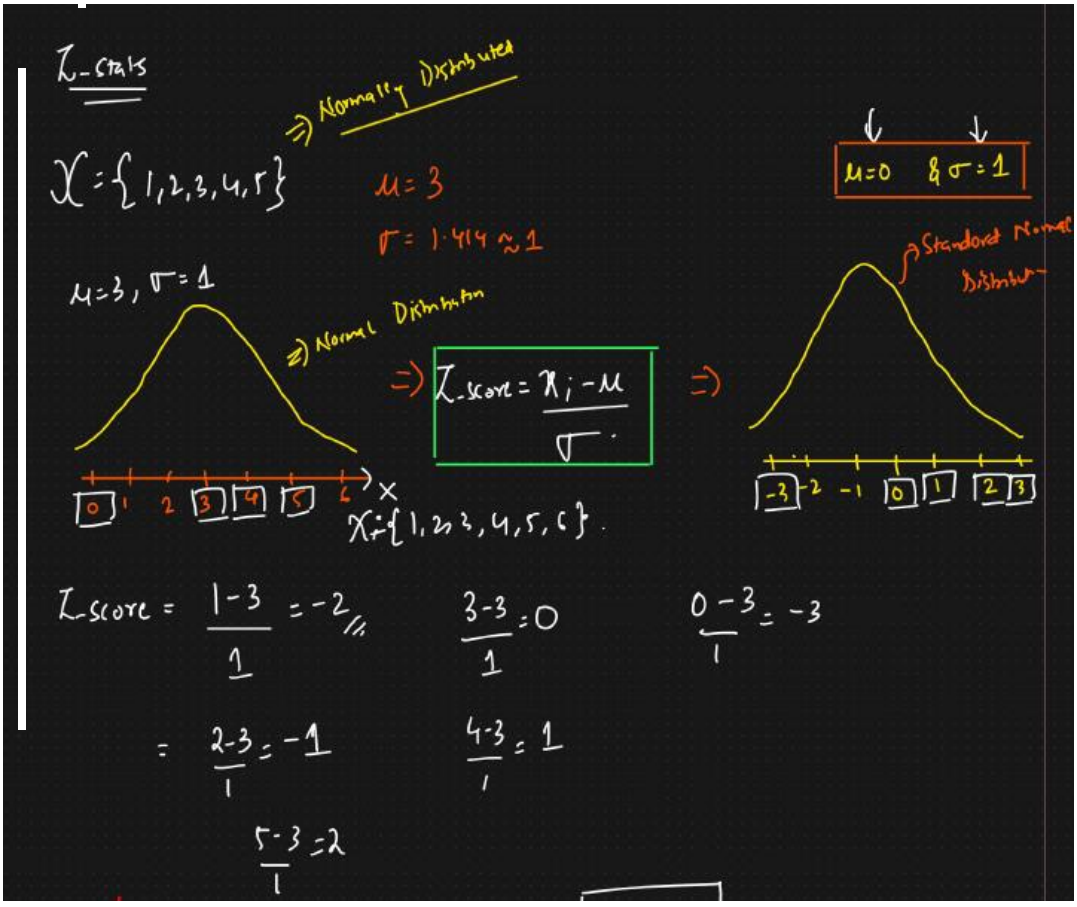
- About 68% of data falls within one standard deviation of the mean
- About 95% of data falls within two standard deviations of the mean
- About 99.7% of data falls within three standard deviations of the mean

- **What is a "Z-score"?**

The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "Z-score". Simply, a Z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units.

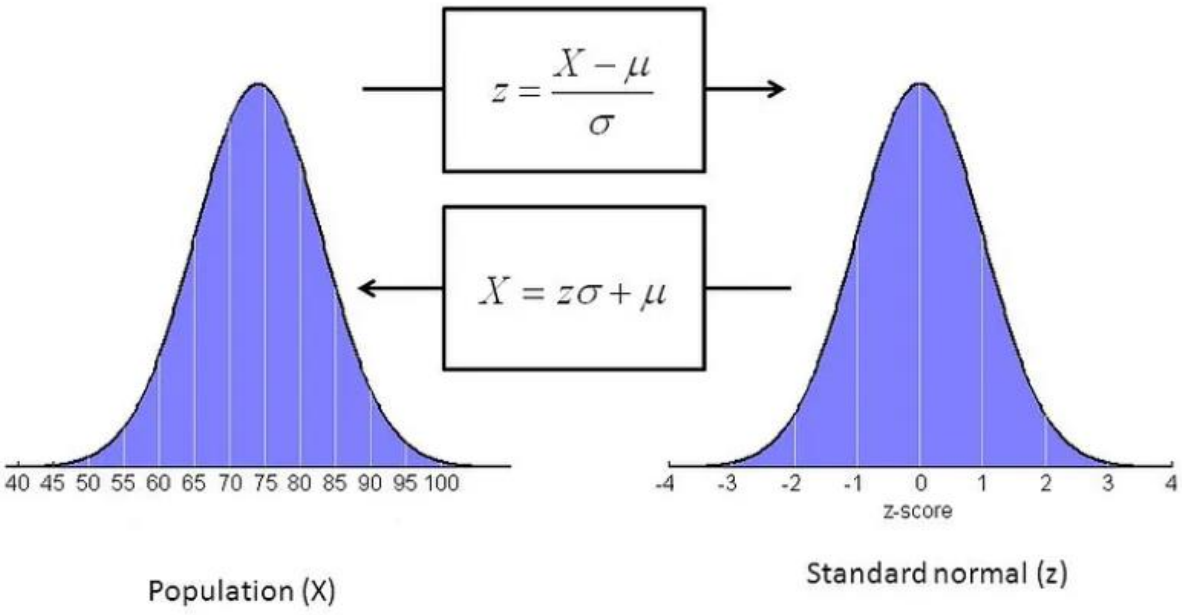
z = (x – μ) / σ

- **Z** is the "z-score" (Standard Score)
- **x** is the value to be standardized
- **μ** (mu) is the mean
- **σ** (sigma) is the standard deviation



Standardizing: - Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

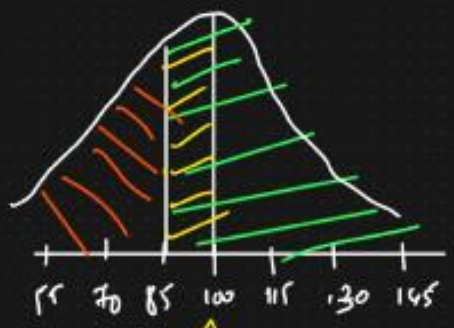
We can take any Normal Distribution and convert it to The Standard Normal Distribution.



S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Prob) In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have and IQ lower than 85?

Ans) $\mu = 100 \quad \sigma = 15$



① Z-score $= \frac{x_i - \mu}{\sigma}$

$= \frac{85 - 100}{15}$

$= -1$

Refer Z table

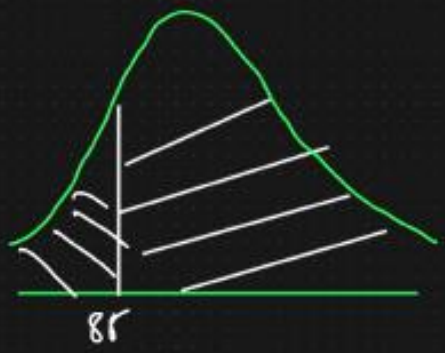
Area under the curve = 0.15866

$0.5 - 0.15866$

$= 0.34134$.

Area under the curve $(> 85) = 1 - 0.15866$

$= 0.8413..$



① How many standard deviation 4.5 is away from mean? ⇒ 0.5

② What percentage of data if falling above 4.5?

Z table ⇒ Area Under the Curve

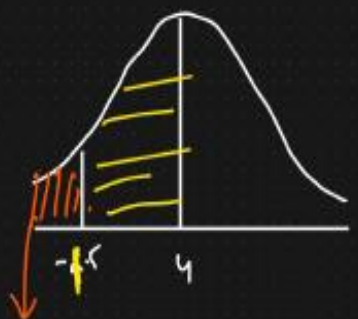
$$Z\text{-score} = \frac{4.5 - 4}{1} = 0.5$$

0.6915

Area under the curve $(\geq 4.5) = 1 - 0.6915 = 0.3085 = 30.85\%$

③ Percentage of data falling below 2.5?

$$Z\text{-score} = \frac{2.5 - 4}{1} = -1.5$$



6.6%

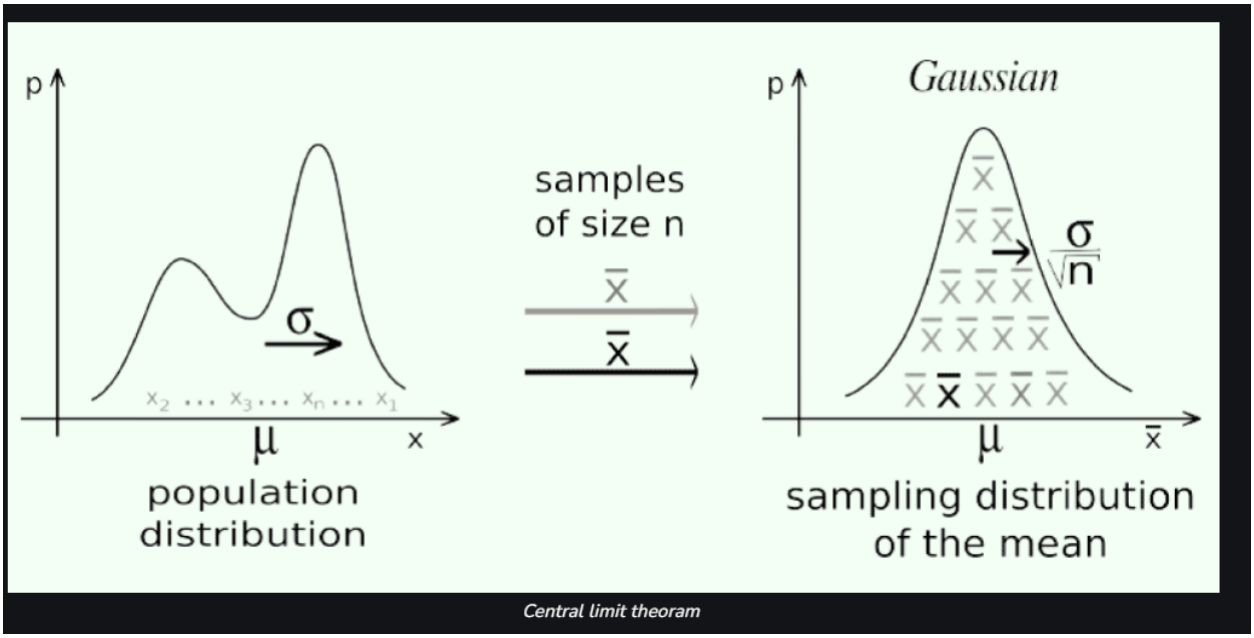
$Z_{table} = 0.0668$

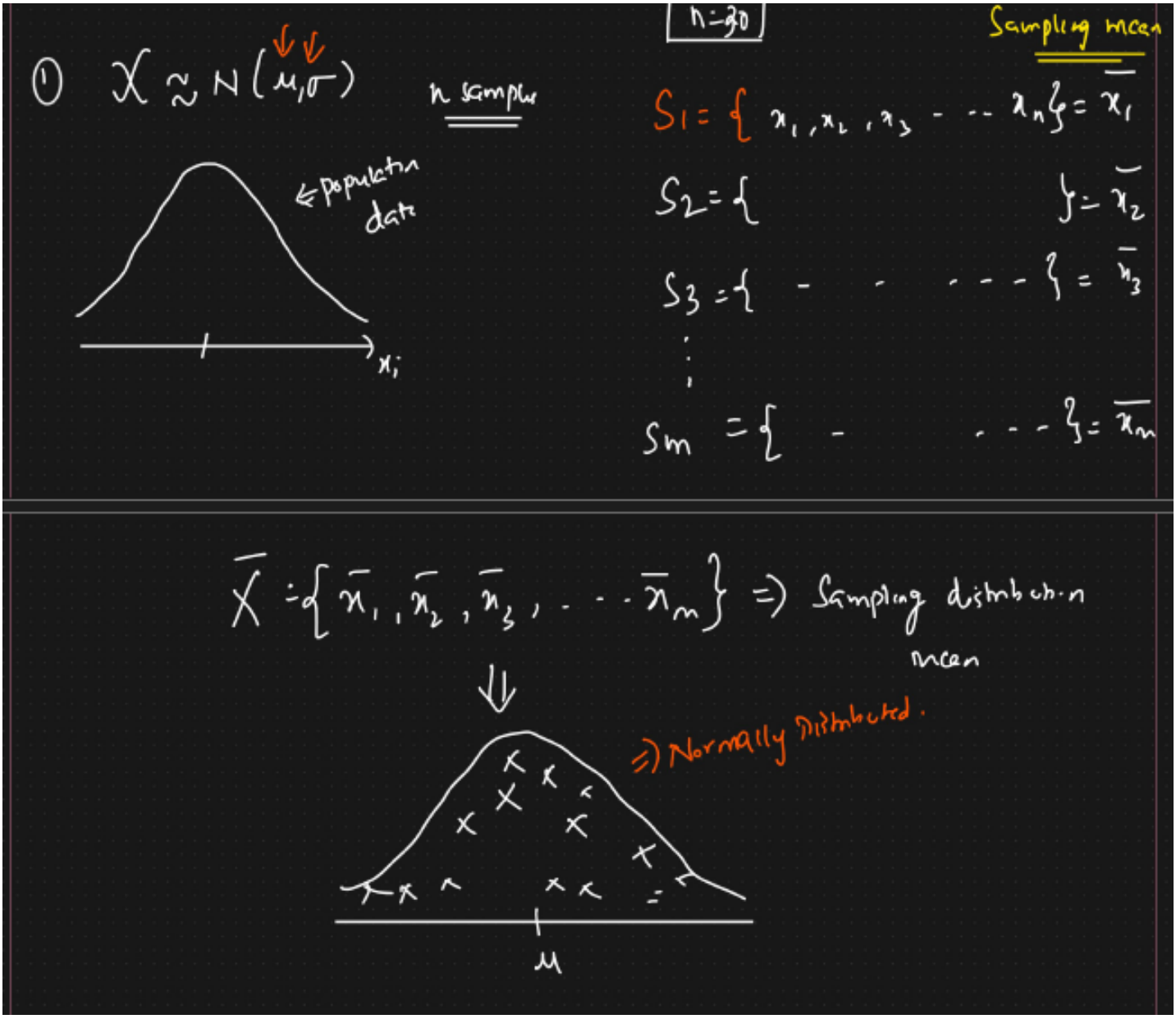
$= 6.6\%$

$0.5 - 0.0668$

Central limit Theorem: - For large sample sizes, the sampling distribution of means will approximate to normal distribution even if the population distribution is not normal.

1. The sample size is **sufficiently large**. This condition is usually met if the size of the sample is $n \geq 30$.
2. The samples are **independent and identically distributed, i.e., random variables**. The sampling should be random.
3. The population's distribution has a **finite variance**. The central limit theorem doesn't apply to distributions with infinite variance.





Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

Sample Standard Deviation = $\frac{\text{Standard Deviation}}{n}$

OR

Sample Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

Notation:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means.
- $\sim$ means "follows the distribution."
- N is the normal distribution.
- μ is the mean of the population.
- σ is the standard deviation of the population.
- n is the sample size.

1. **What is Central Limit Theorem in Statistics?**

Central Limit Theorem in statistics states that whenever we take a large sample size of a population then the distribution of sample mean approximates to the normal distribution.

2. **When does Central Limit Theorem apply?**

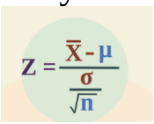
Central Limit theorem applies when the sample size is larger usually greater than 30.

3. **Why is Central Limit Theorem important?**

Central Limit Theorem is important as it helps to make accurate prediction about a population just by analyzing the sample.

4. **How to solve Central Limit Theorem?**

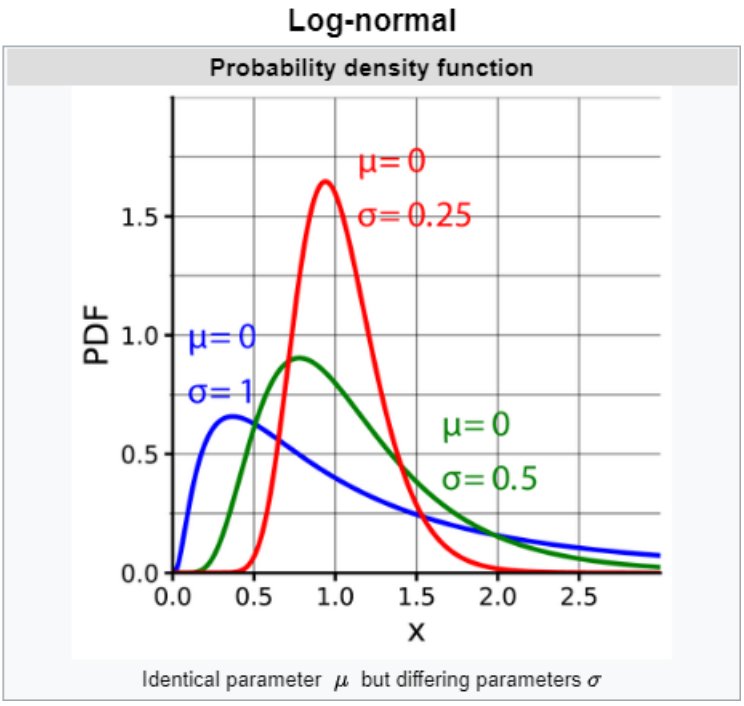
The Central Limit Theorem can be solved by finding Z score which is calculated by using the formula.

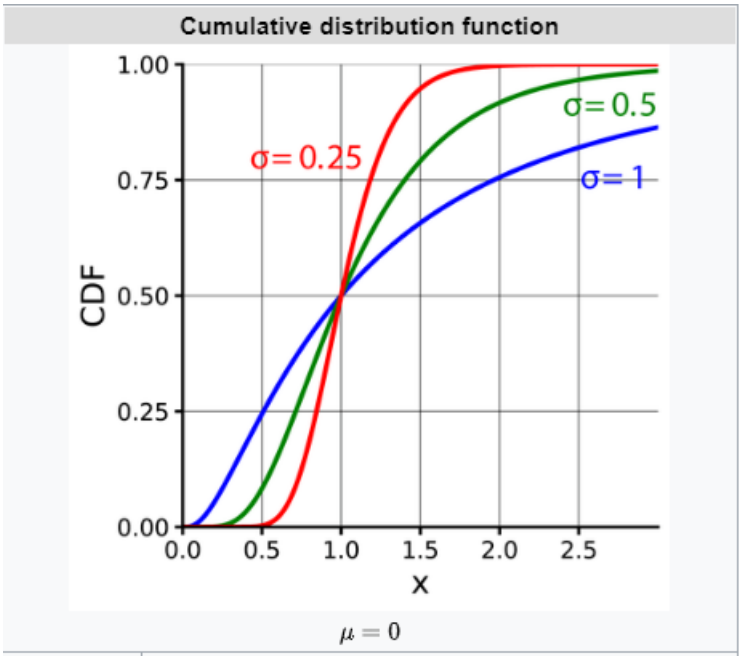


how to check if distribution is normal or not

If you want to check the normal distribution using a histogram, plot the normal distribution on the histogram of your data and check that the distribution curve of the data approximately matches the normal distribution curve. A better way to do this is to use a quantile-quantile plot, or Q-Q plot for short.

6. **Log-Normal Distribution: -** A log-normal distribution is a continuous distribution of random variable y whose natural logarithm is normally distributed. For example, if random variable y = exp { y } has log-normal distribution then x = log (y) has normal distribution.



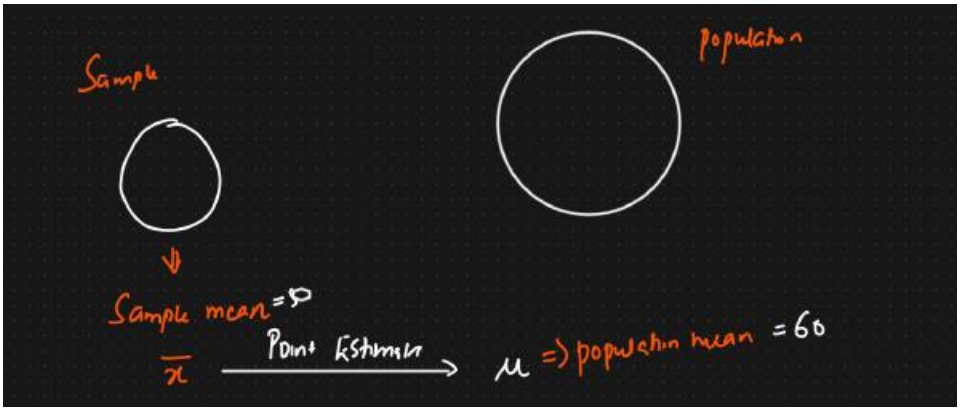


➢ Inferential Statistics

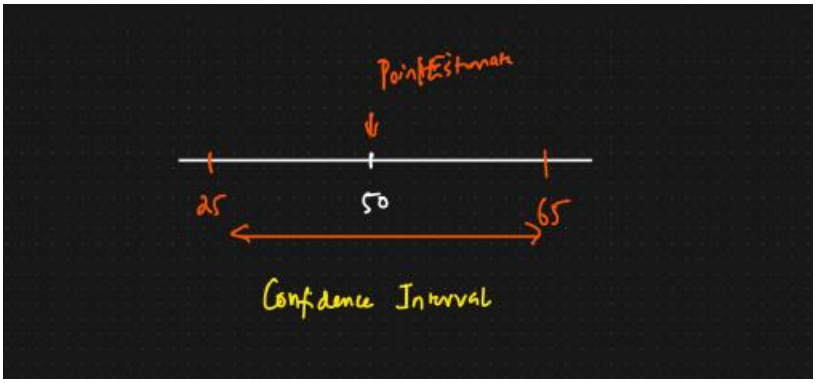
Statistical inference provides methods for drawing conclusions about a population from sample data.

1. **Estimate: -** it is an observed numerical value used to estimate an unknown population parameter

   I. **Point Estimate**: - Single numerical value used to estimate the unknown population parameter.



   II. **Interval Estimate**: - Range of value used to estimate the unknown Population Parameter

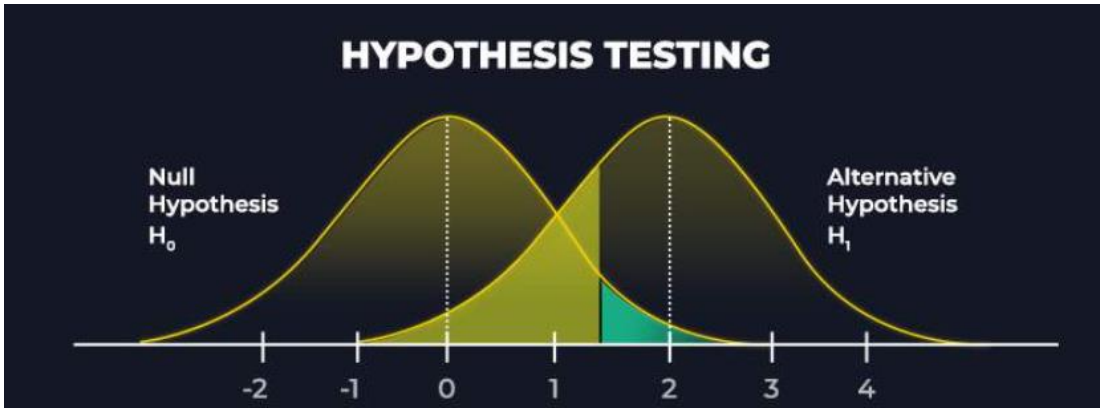


2. Hypothesis And Hypothesis Testing Mechanism: -

Inferential Stats is a Conclusion or inferences about the population data

- **Hypothesis Testing Mechanism: -** Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution

  - Null Hypothesis (H0):- The Null Hypothesis (H0) aims to nullify the alternative hypothesis by implying that there exists no relation between two variables in statistics. It states that the effect of one variable on the other is solely due to chance and no empirical cause lies behind it.
  - Alternative Hypothesis (H1):- Alternative Hypothesis (H1) or the research hypothesis states that there is a relationship between two variables (where one variable affects the other). The alternative hypothesis is the main driving force for hypothesis testing.



Hypothesis Testing Mechanisms

Person Crime →

① Null Hypothesis (H_0) = The person is not guilty

- Assumption you are beginning with

② Alternate Hypothesis (H_1) = The person is guilty

- Opposite of Null Hypothesis

Z test, Chi square test, t test, ANOVA.

③ Evidences → (DNA, Finger Test, - ....) ⇒ Experiments ⇒ Statistical Analysis

④ We fail to reject the Null Hypothesis or Reject the Null Hypothesis

↓ person is not guilty

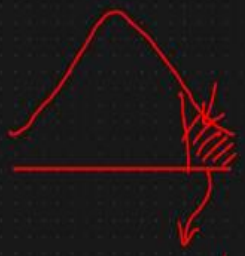
↓ person is guilty

Eg: Colleges at District A states its average placed percentage of students are 85%. A new college opened in the district and it was found Sample of student 100 have a pass percentage of 90% with a Standard deviation of 4%. Does this School have a different passed percentage?

>85%

Ans) Null Hypothesis H_0 = $\mu = 85\%$

Alternate Hypothesis H_1 $\mu \neq 85\%$ $\mu > 85\%$

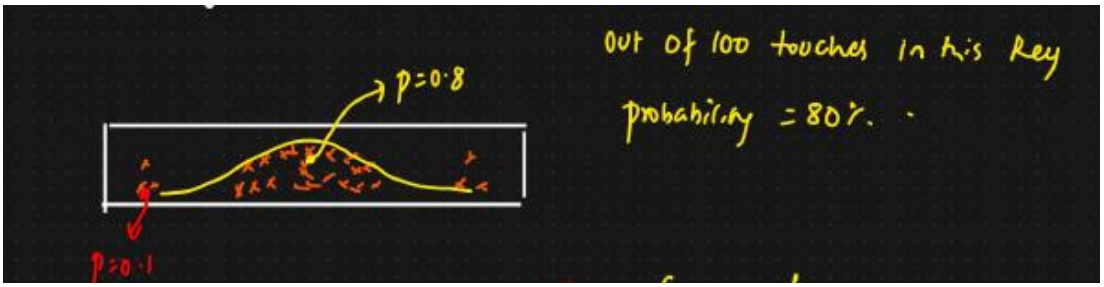


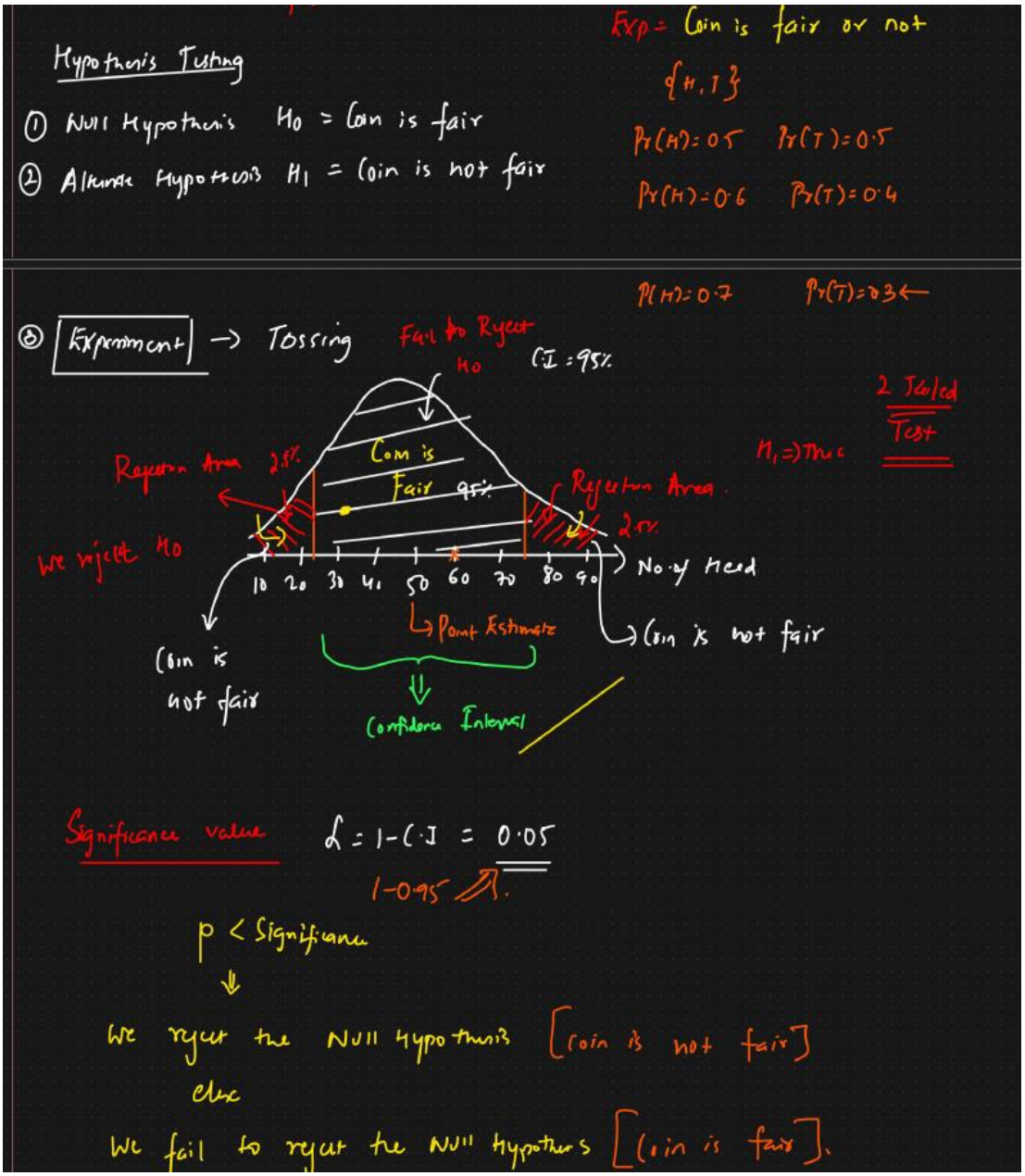
one tailed Test

Does this school have a passed percentage < 85%

H_1 : $\mu < 85\%$

3. **P-Value: -** P value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observation if the null hypothesis were true, p values are used in hypothesis testing to help decide whether to reject the null hypothesis



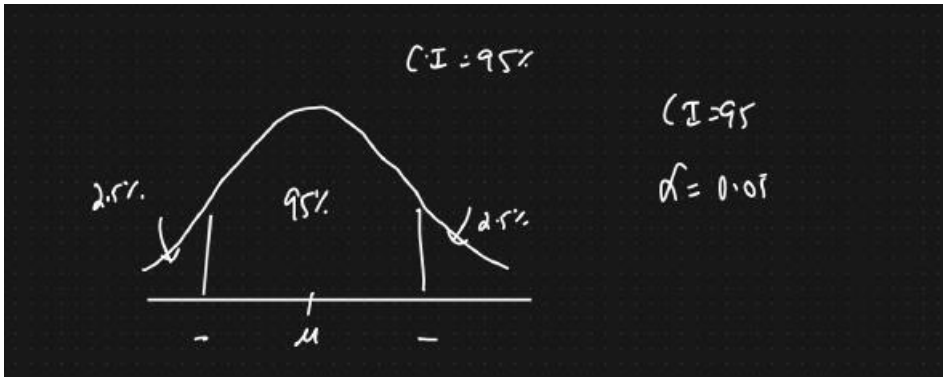


4. **Confidence Interval and Margin of Error: -** Confidence intervals are a range of values within which we can be confident that the true population parameter lies. This range is estimated based on a sample from the population and a chosen level of confidence. The level of confidence speaks to the likelihood that the genuine populace parameter lies inside the certainty interim.

   Confidence Interval = [lower bound, upper bound]

   The **margin of error** is equal to half the width of the entire confidence interval.

   lower bound, upper bound = sample mean ± margin of error



Margin of Error

- In order to find a confidence interval, the margin of error must be known.
- The margin of error depends on the degree of confidence that is required for the estimation.
- Typically degrees of confidence vary between 90% and 99.9%, but it is up to the researcher to decide.

$$\text{Margin of error} = z^* \cdot \frac{\text{population standard deviation}}{\sqrt{n}}$$

- The level of confidence is represented by z* (called z star).
- It is also necessary to know the standard deviation of the variable in the population. (Note: the population standard deviation is NOT the same as the sample standard deviation).
- Finally, the size of the sample n will be used to compute the margin of error.

Confidence Intervals

- The formulas for the confidence interval and margin of error can be combined into one formula.

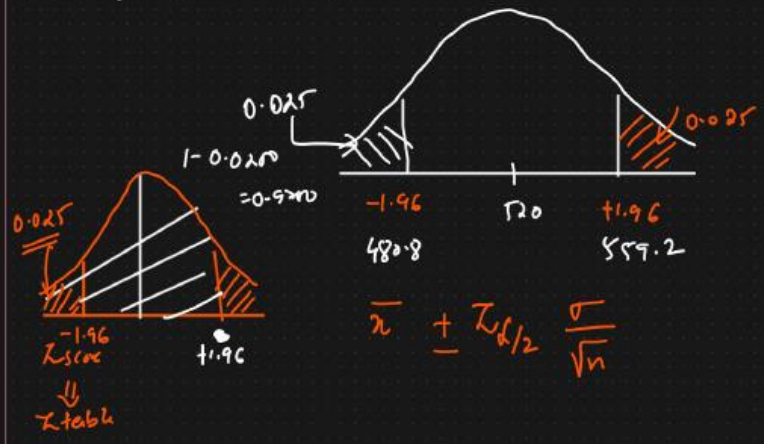
Confidence interval = sample mean ± margin of error

Confidence Interval Formula:

Confidence interval= sample mean ± z* x Population standard deviation/√N

① In the verbal section of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% C.I about the mean?

Ans) $\sigma = 100$ $n = 25$ $\bar{x} = 520$ C.I = 0.95 $\alpha = 0.05$



$Z_{0.05/2} = Z_{0.025} =$

$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Lower C.I = $520 - (1.96) * \frac{100}{\sqrt{25}} = 480.8$

Higher C.I = $520 + (1.96) * \frac{100}{\sqrt{25}} = 559.2$

I am 95% Confident (or sure) that the mean CAT score lies between 480.8 and 559.2.

➢ **Hypothesis Testing and Statistical Analysis: -**

1. **Z-Test** } **Average**
2. **T-Test** }
3. **Chi Square ---------→ Categorical**
4. **Anova--------→ Variance**

1. Z-Test:-

- Population standard deviation is known
- Large sample size (n > 30)
- **Z-Test** = $(\bar{x} - \mu) / (\sigma / \sqrt{n})$
  $\sigma/\sqrt{n}$----→ Standard Error
  σ -----→ Population standard deviation
  μ-----→ Population Mean
  $\bar{x}$-----→ Sample Mean
  n----→ No. of Sample
- Degrees of Freedom Not applicable
- We Used Z Test when the population standard deviation is known and the sample size is large

The z-test is also a hypothesis test in which the z-statistic follows a normal distribution. The z-test is best used for greater-than-30 samples because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed.

Confidence interval = Point Estimate ± margin of error
Confidence interval = sample mean ± margin of error
C.I=$\bar{x} \pm Z_{\alpha/2} * \sigma/\sqrt{n}$
$\sigma/\sqrt{n}$----→ Standard Error
σ -----→ Population standard deviation
α -----→significance level
n-----→ no. of samples

① The average heights of all residents in a city is 168cm with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5cm.

① State Null And Alternate Hypothesis

② At a 95% Confidence level, is there enough evidence to Reject Null Hypothesis

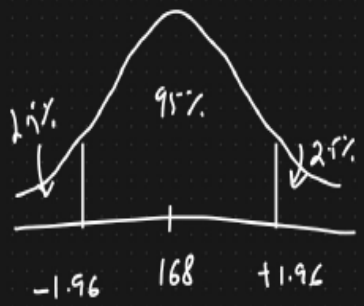
Ans) $\mu = 168cm$ $\sigma = 3.9$ $n = 36$ $\bar{x} = 169.5cm$.

a) Null Hypothesis H_0 $\mu = 168cm$

b) Alternate Hypothesis H_1 $\mu \neq 168cm$

c) $C.I = 0.95$ $\alpha = 1 - 0.95 = 0.05$

d) Decision Boundary

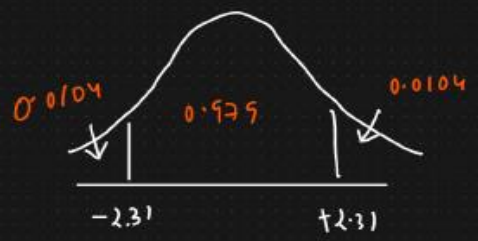


e) Statistical Analysis

$$Z_{test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.91/\sqrt{36}} = 2.31$$

Conclusion $Z_{test} > 1.96$ { We Reject H_0 }.

f) P-value



P-value = 0.0104 + 0.0104 = 0.0208

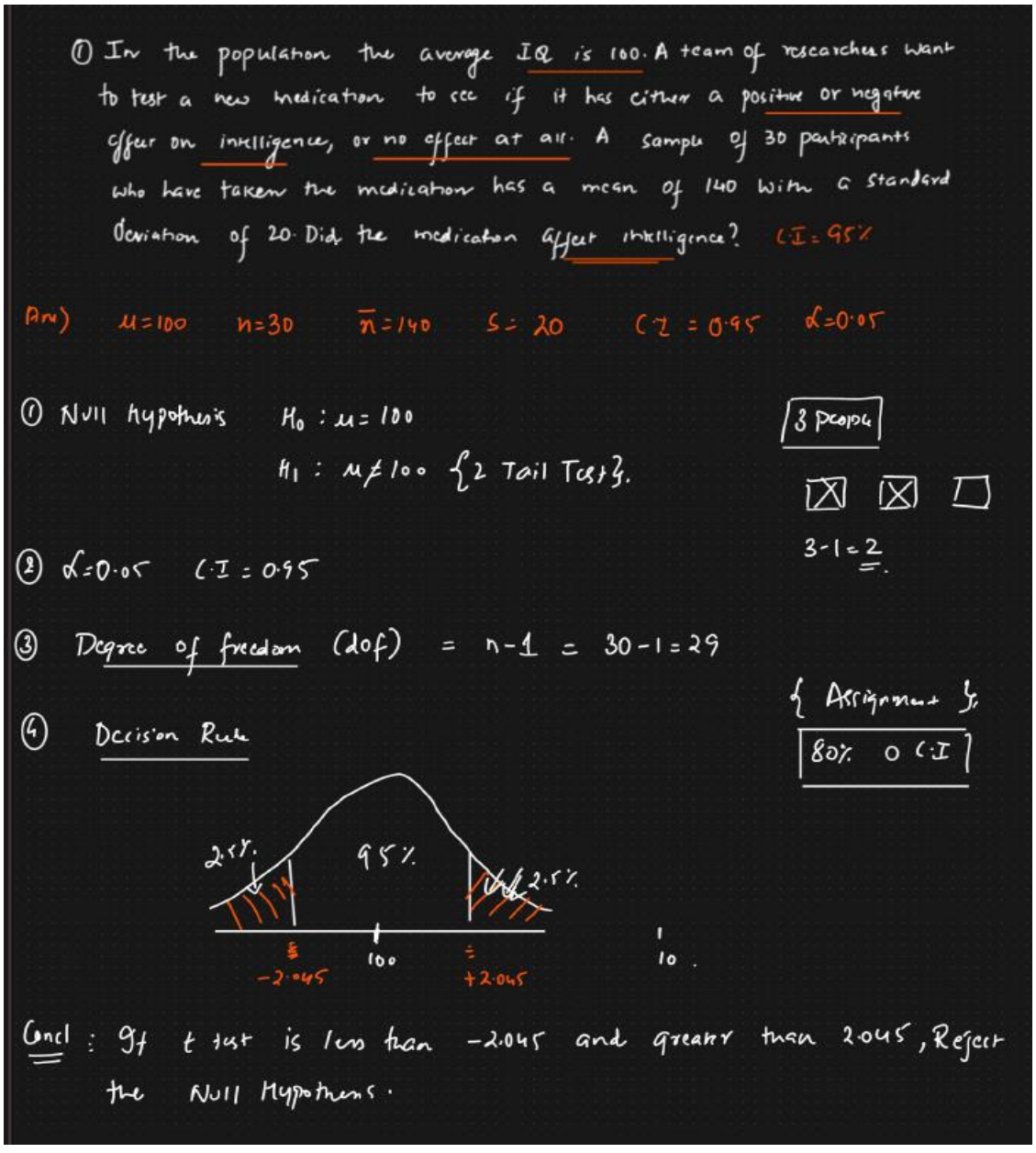
If p-value < Significance

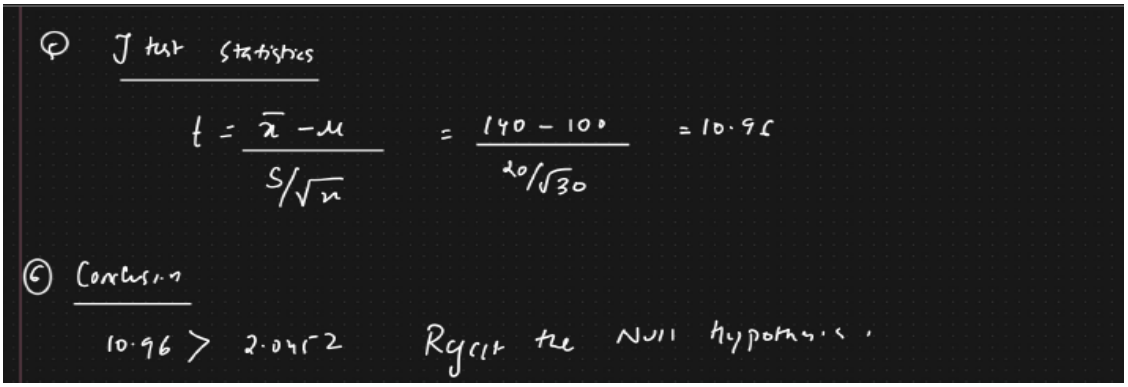
0.0208 < 0.05 { We Reject Null Hypothesis }.

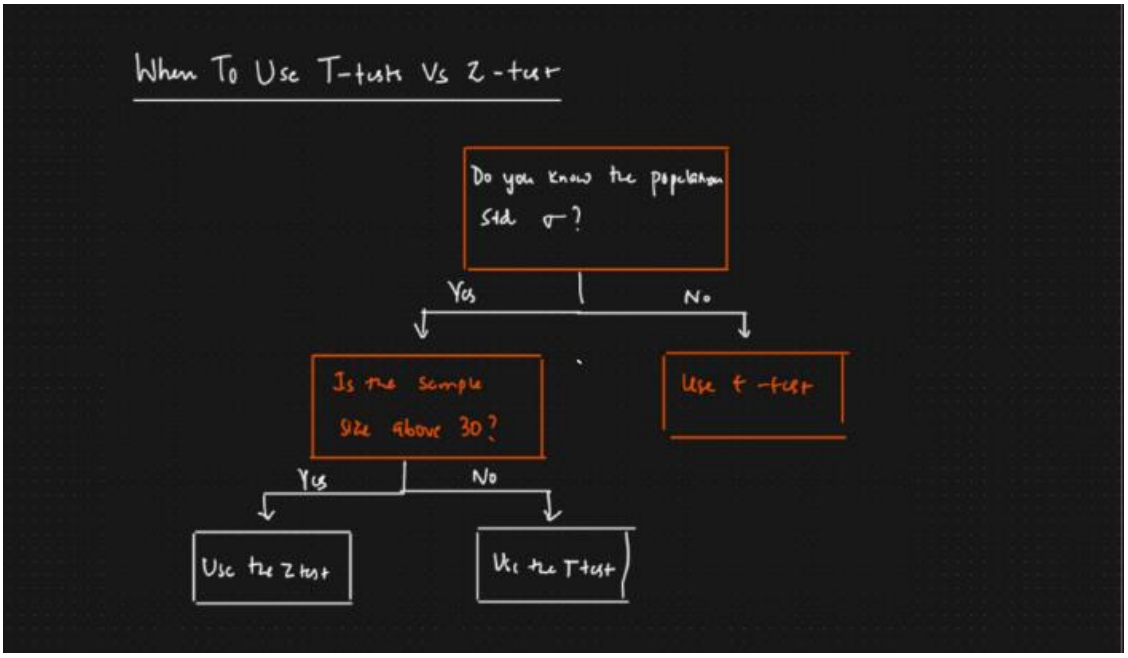
2. **T-Test: -** A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.
    - Population standard deviation is unknown
    - Our sample size is small, n < 30
    - T-Test = $(\bar{x} - \mu) / (s / \sqrt{n})$
      $\sigma/\sqrt{n}$----→ Standard Error
      s -----→ sample standard deviation
      μ-----→ Population Mean
      $\bar{x}$-----→ Sample Mean
      n----→ No. of Sample

- Degrees of Freedom is n-1
- We Used T-Test when the population standard deviation is unknown or the sample size is small
- T-tests can be dependent or independent.

Confidence interval = Point Estimate ± margin of error
Confidence interval = sample mean ± margin of error
C.I=$\bar{X}$ ± T $_{\alpha/2}$* s/√n
s/√n------→ Standard error
s-----→ Sample variance
α -----→significance level
n-----→ no. of samples







- Z-Test & T-Tests are Parametric Tests, where the Null Hypothesis is less than, greater than or equal to some value.
- A z-test is used if the population variance is known, or if the sample size is larger than 30, for an unknown population variance.
- If the sample size is less than 30 and the population variance is unknown, we must use a t-test.

Q1. When Are Z-test and T-test Used?

A. A z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

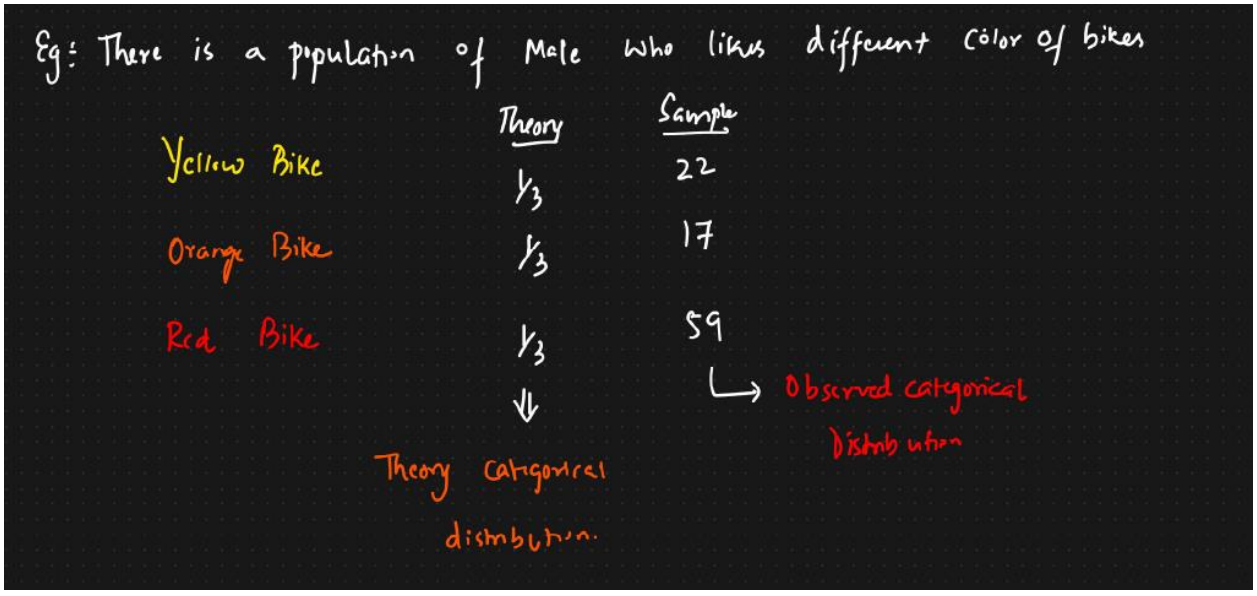
Q2. What Is the Difference Between a Two-Tailed and One-Tailed Z-Test?

A. A one-tailed z-test allows for the possibility of rejection of the Null Hypothesis in only one direction, whereas a two-tailed z-test tests the possibility of rejection in both directions (left and right).

Q3. What Are the Assumptions of the T-Test and Z-Test?

A. It is assumed that the z-statistic follows a standard normal distribution, whereas the t-statistic follows the t-distribution with a degree of freedom equal to n-1, where n is the sample size

3. **Chi Square: -**
    - Chi Square test clams about Population proportions
    - It is a non-parametric test is performed on categorical (nominal or ordinal) data



Goodness of fit

(x) In a student class of 100 students, 30 are Right handed. Does this class fit the theory 12% of people are right handed.

	O	E
Right handed	30	12
left handed	70	88

⇒ Theory Categorical Distribution

⇓

Observed Categorical Distribution

Chi Square For Goodness of fit

In 2010 Census of the city, the weight of the individuals in a small city were found to be the following

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are the results

<50	50-75	>75
140	160	200

CI = 95%

⇑

Using $\alpha = 0.05$, would you conclude the population differences of weights has changed in the last 10 years?

Ans)

2010

<50kg	50-75 kg	>75
20%	30%	50%

Expected

$\frac{20}{100} \times 500$

$\frac{30}{100} \times 500$

$\frac{50}{100} \times 500$

n=500

Observed

2020

<50	50-75	>75
140	160	200

Expected

<50	50-75	>75
100	150	250

① Null Hypothesis : H_0 : The data meets the expectation

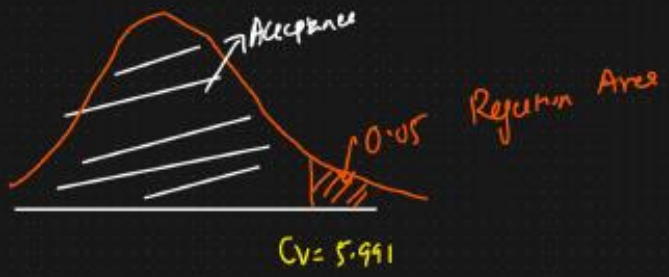
② Alternate Hypothesis : H_1 : The data does not meet the expectation

③ $\alpha = 0.05$ C.I = 95%

④ Degree of freedom

$df = K - 1 = 3 - 1 = 2$

⑤ Decision Boundary → CHI Square Table



CV = 5.991

If χ^2 is greater than 5.991, Reject Ho

use

We fail to Reject Ho.

⑤ Calculate Chi Square Test Statistics

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Observed

n = 500

2020

<50	50-75	>75
140	160	200

Expected

<50	50-75	>75
100	150	250

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$= 16 + 0.66 + 10$$

$$\chi^2 = 26.66$$

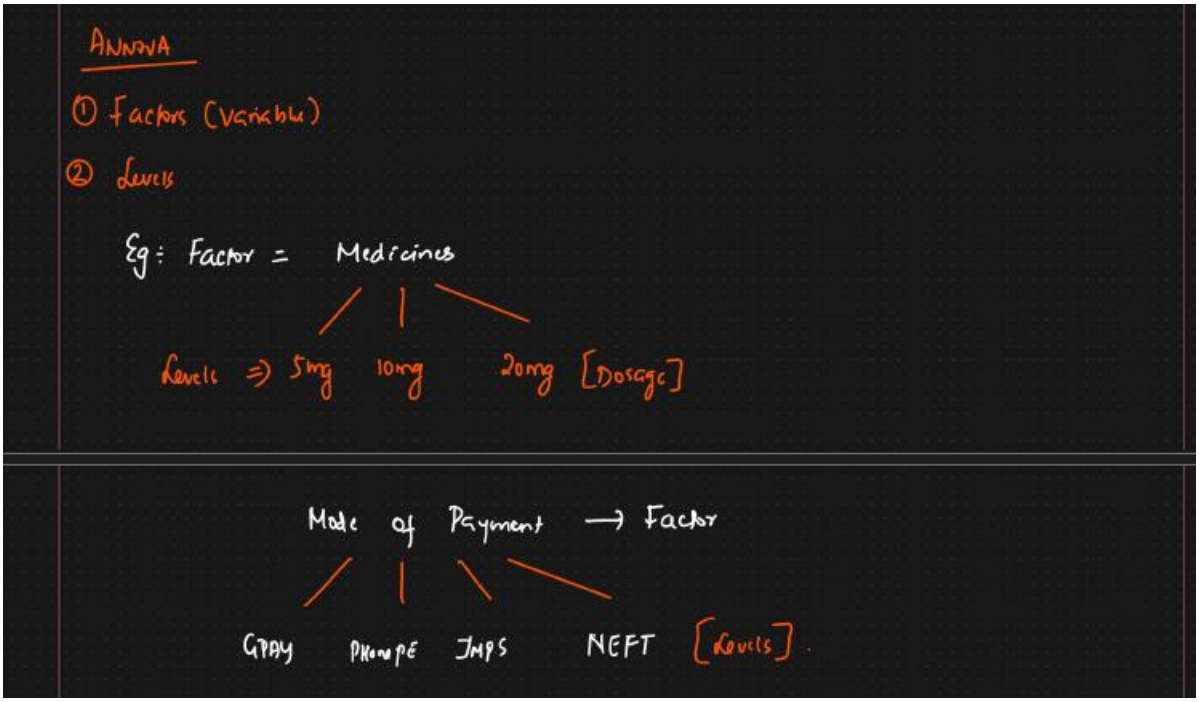
Assignment

$\alpha = 0.25$

26.66 > 5.99, Reject Ho

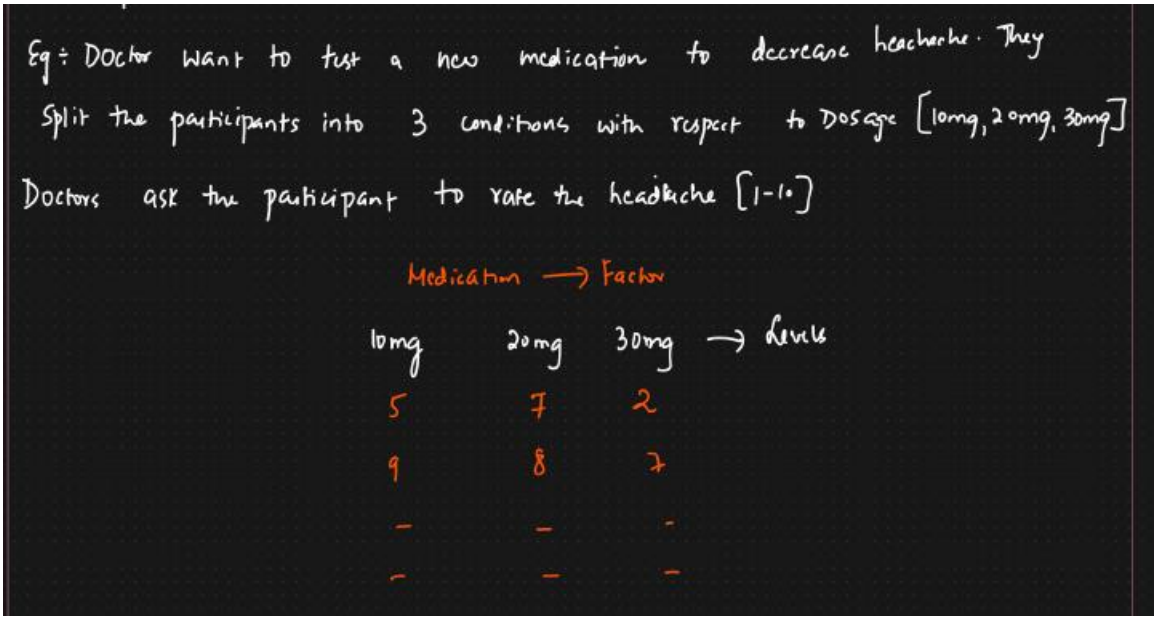
4. Anova(F-Test): -

- **ANOVA**, which stands for Analysis of Variance, is a statistical test used to analyze the difference between the means of more than two groups.
- ANOVA compares the variation between group means to the variation within the groups. If the variation between group means is significantly larger than the variation within groups, it suggests a significant difference between the means of the groups.
- ANOVA calculates an F-statistic by comparing between-group variability to within-group variability. If the F-statistic exceeds a critical value, it indicates significant differences between group means.
- ANOVA is used to compare treatments, analyse factors impact on a variable, or compare means across multiple groups.
- Types of ANOVA include one-way (for comparing means of groups) and two-way (for examining effects of two independent variables on a dependent variable).

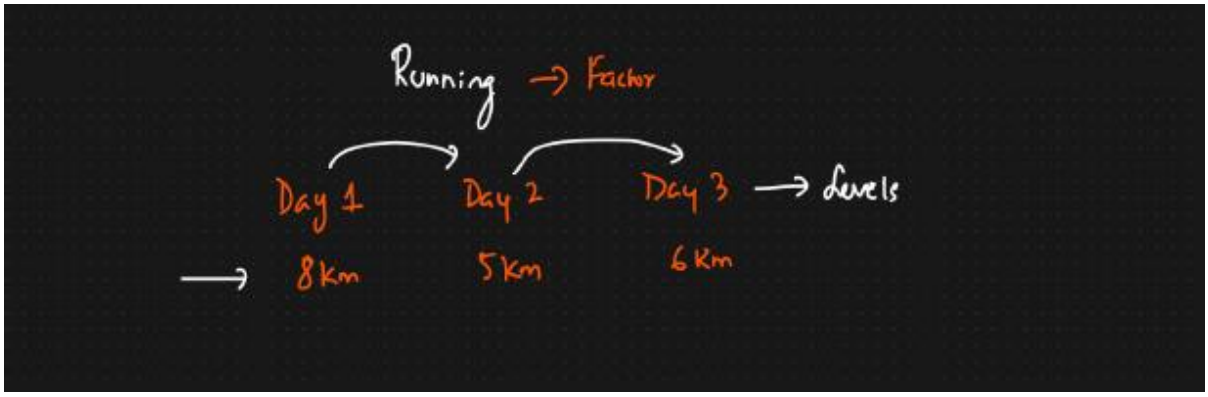


Types of Anova

1. **One Way Annova:- One factor with at least 2 levels, these levels are independent**

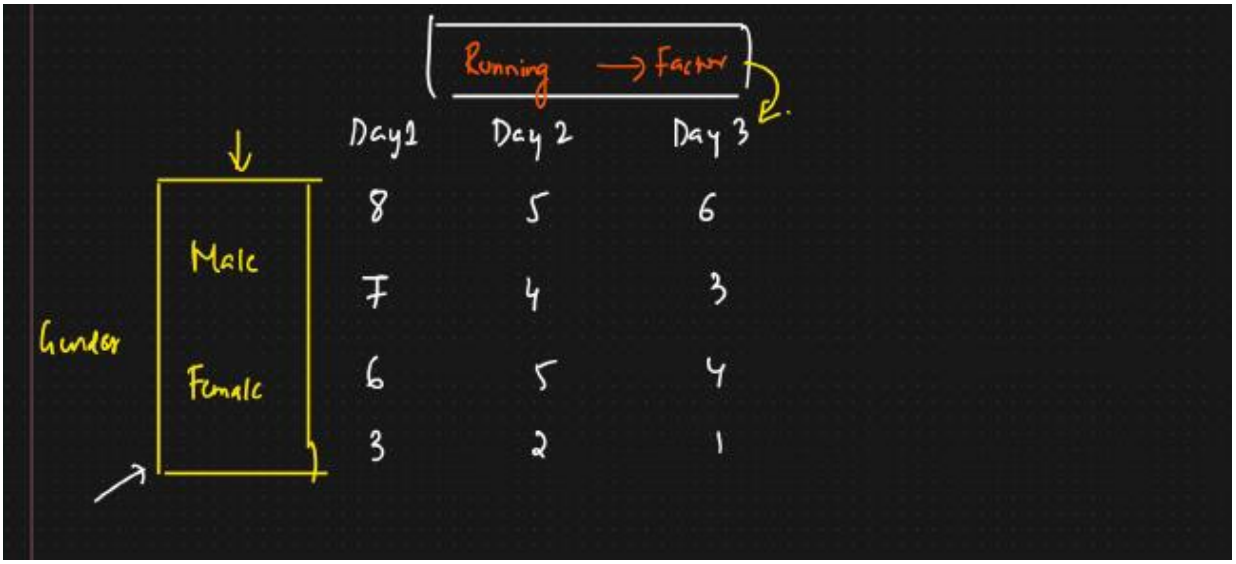


2. **Repeated measures annova:-** One factor with atleast 2 levels, levels are dependents



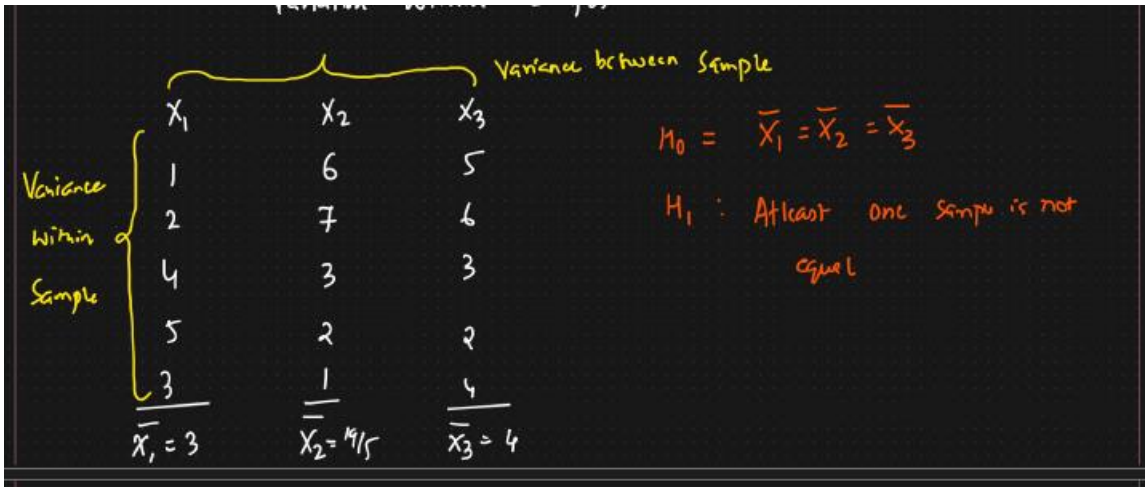
3. Factorial Annova:- Two or More factors (Each of which with at least 2 levels)

Levels can be either independent or dependent



- Hypothesis Testing of Annova:-
  - Null Hypothesis H_0 : $\mu_1 = \mu_2 = \mu_3 = - - - - - \mu_k$
  - Alternate hypothesis H_1 : At least one of mean is not equal
  - F Test Statistics

F = Variation between Samples / variation within samples



➢ **One Way Annova:- One Factor with at least 2 levels, levels are independent**

① Doctors want to test a new medication which reduces headache. They split the participant into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10]. Are there any differences between the 3 conditions using alpha = 0.05 ?

15 mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① Define Null Hypothesis H_0: $\mu_{15} = \mu_{30} = \mu_{45}$

② Alternate " " H_1: Atleast one mean is not equal

③ State Significance Value

$\alpha = 0.05 \Rightarrow CI = 0.95$

④ Calculate Degree of freedom

$N = 21 \quad a = 3 \quad n = 7$

$df_{between} = a - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

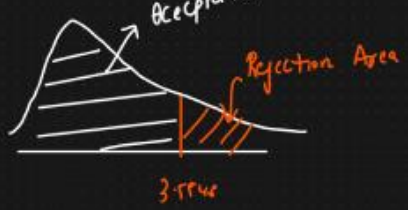
$\downarrow$

$(2, 18) \Rightarrow$ F Test Table

$df_{total} = N - 1 = 21 - 1 = 20$

Decision Boundary

$\alpha = 0.05$



If F test is greater than 3.5546, Reject the Ho.

⑤ Calculate F Test Statistics

	SS	df	MS	F
Between	98.67	2	49.34	86.56
Within	10.29	18	0.57	
Total	108.95	20		

15 mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① $SS_{between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

15mg = 9+8+7+8+8+9+8 = 57

30mg = 7+6+6+7+8+7+6 = 47

45mg = 4+3+2+3+4+3+2 = 21

$\Rightarrow T^2$

$= \frac{57^2+47^2+21^2}{7} - \frac{[57+47+21]}{21}$

$= 98.67$

$\sum y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2$
$+7^2+6^2+6^2+7^2+8^2+7^2+6^2$
$+ \cdots$
$= 853$

② $SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$= \Sigma y^2 - \left[\frac{57^2 + 47^2 + 21^2}{7}\right]$

$= 853 - \left[\frac{57^2 + 47^2 + 21^2}{7}\right]$

$= \boxed{10.29}$

③ $SS_{total} = \Sigma y^2 - \frac{T^2}{N}$

$= 853 - \frac{125^2}{21} = \boxed{108.95}$

$F = \frac{\text{Variation between samples}}{\text{Variation within samples}} = \frac{MS\ Between}{MS\ within}$

$= \frac{49.34}{0.54} = 86.56$ //.

$86.56 > 3.556$, Reject the Null Hypothesis