

Business Statistics



Agenda

01 Descriptive Statistics

- Measure of central tendency
- Measure Of Dispersion

02 Probability

- Probability Distributions
-

03 Inferential Statistics

- Type of sampling Technique
- Central limit theorem
- Confidence interval

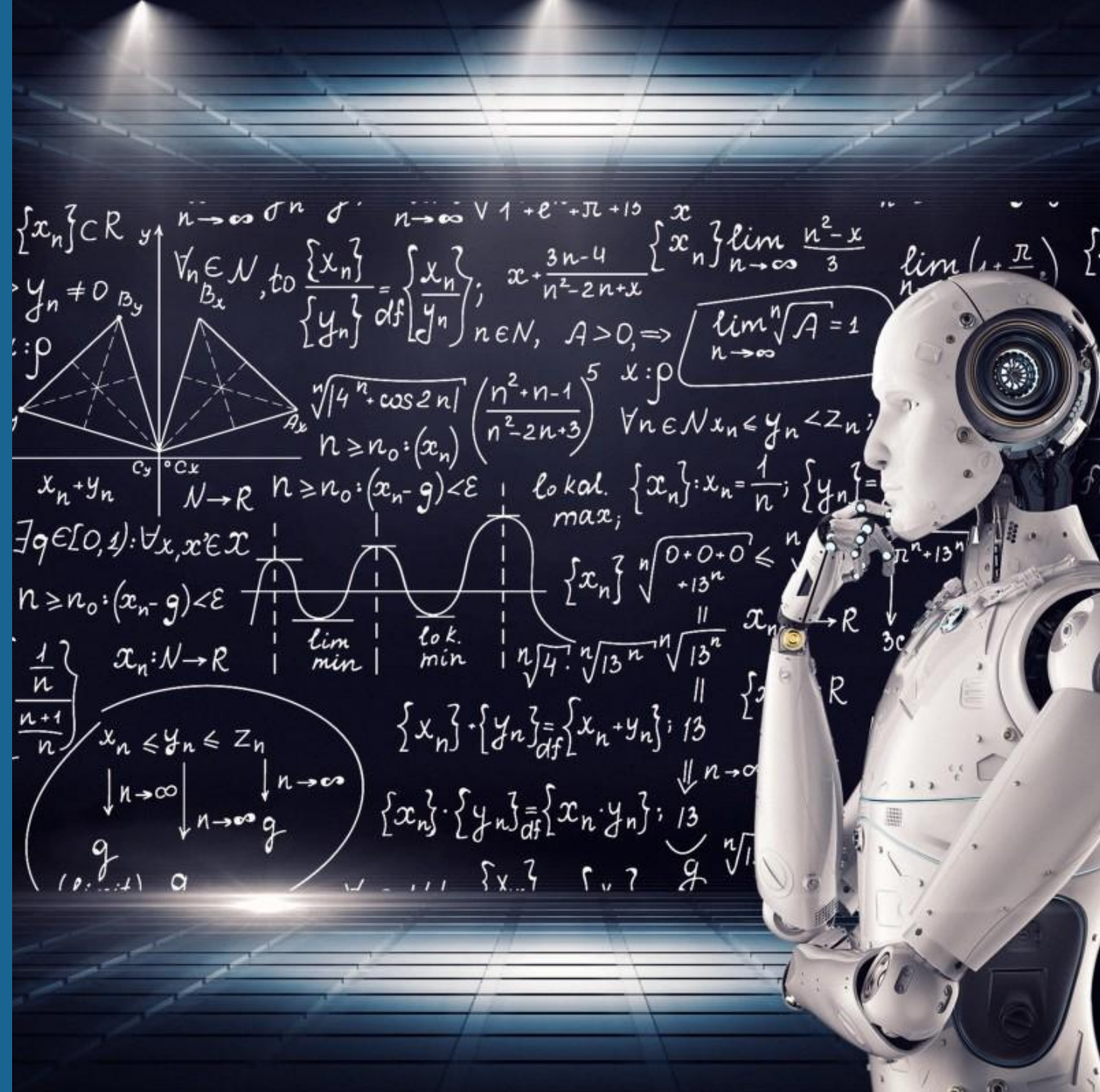
04 Hypothesis testing

- Significance level
- P-value
- Statistical tests

Statistics

1. **Statistics** is concerned with the describing, interpretation & analyzing of the data
2. Statistics:
 - Descriptive Statistics
 - Inferential Statistics
3. It uses **analytical methods** which provide the math to model & predict variation
4. It uses **graphical methods** to help making numbers visible for communication purposes.

Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data.



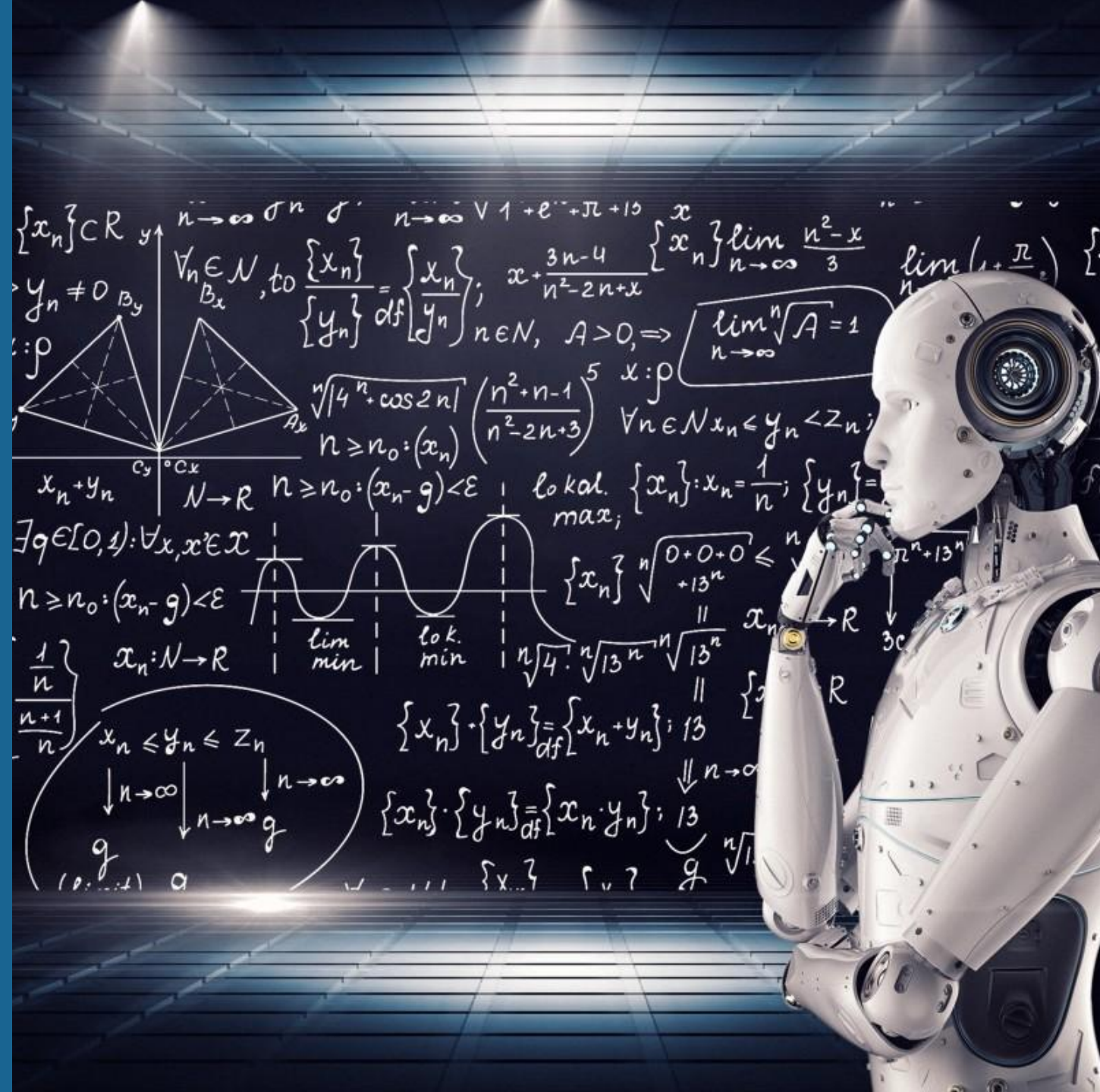
Descriptive Statistics

Descriptive Statistics

Descriptive Statistics describe/summarize the data and provide initial finding for the data. It give us basic understanding for the data. Descriptive statistics doesn't give us any conclusion about the data but give the brief understanding about the data. It will give us mean/average, median, Mode ,Standard deviation etc about the data.

Types of Descriptive Statistics:-

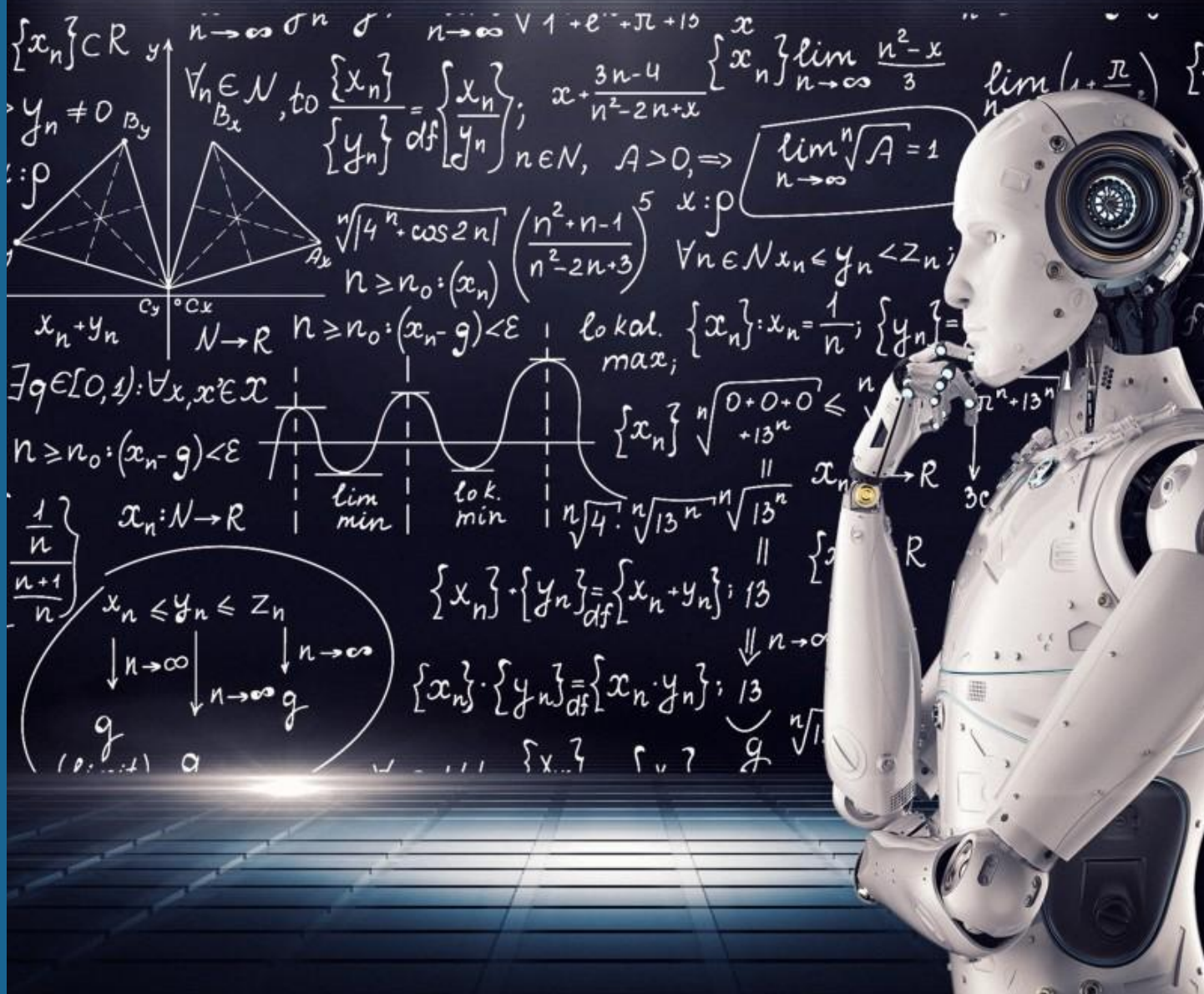
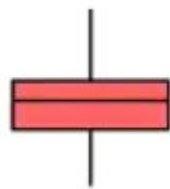
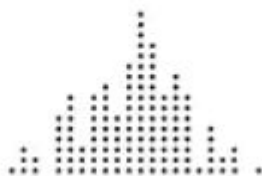
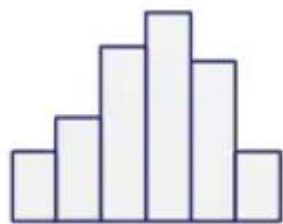
1. Measure of Central Tendency
2. Measure of Dispersion



When analyzing a graphical display, you can draw conclusions based on several characteristics of the graph.

You may ask questions such as:

1. Where is the approximate middle, or centre, of the graph
2. How spread out are the data values on the graph
3. What is the overall shape of the graph
4. Does it have any interesting patterns



Outlier:

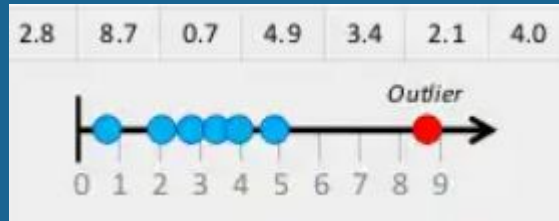
1. A data point that is significantly greater or smaller than other data points in a dataset
2. It is useful when analyzing data to identify outliers
3. They may affect the calculations of descriptive statistics
4. Outliers can occur in any given dataset and in any distribution

The easiest way to detect them is by **graphing the data** or using graphical methods such as:

1. Histograms
 2. Box Plots
 3. Normal distribution plots
- Outliers may indicate an experimental error or incorrect recording of data
 - They may also occur **by chance**
 - It may be normal to have high or low data points
 - You need to decide whether to exclude them before carrying out your analysis
 - An outlier should be excluded if it is due to human error

Outlier:

This example is about the time taken to process a sample of applications



It is clear that one data point is far from the rest of the values, and hence is called as an **Outlier**

Detecting Outliers:

1. Standard Deviation: In statistics, If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations

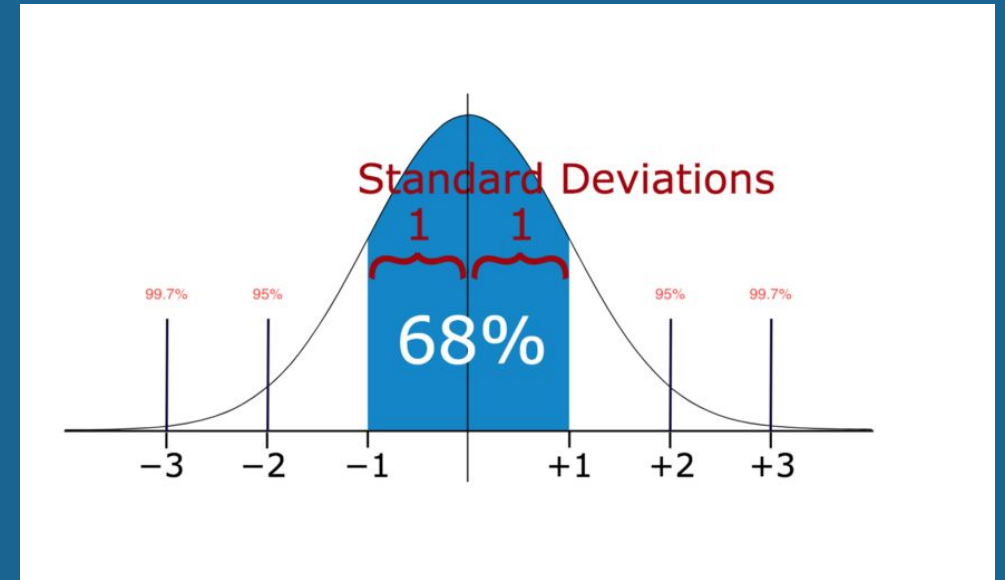
Lets, understand what is standard deviation:

The Standard Deviation is a measure of how spread out numbers are. Its symbol is σ (the greek letter sigma)

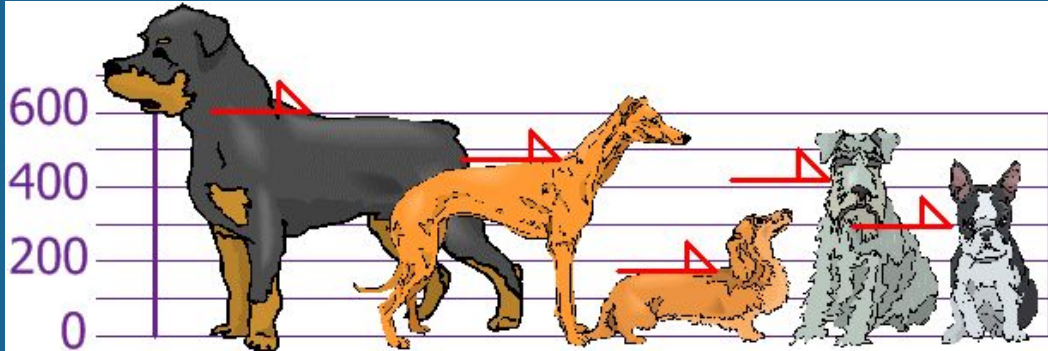
The formula is easy: It is the square root of the Variance.

So now you ask, "What is the Variance?"

Variance: The average of the squared differences from the Mean.

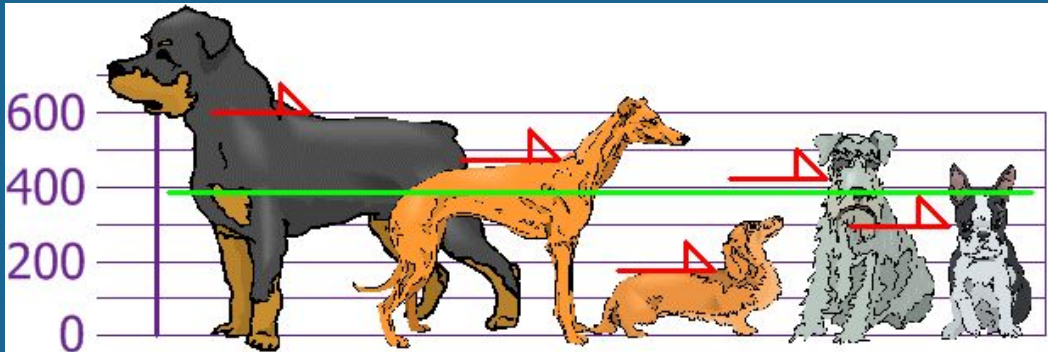


The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm

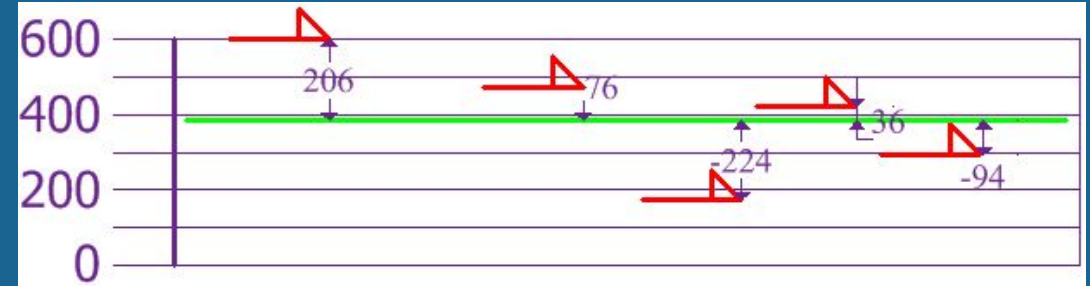


Find out mean, variance & standard deviation

$$\text{Mean} = 600 + 470 + 170 + 430 + 300 / 5 = 394$$



Now we calculate each dog's difference from the Mean:

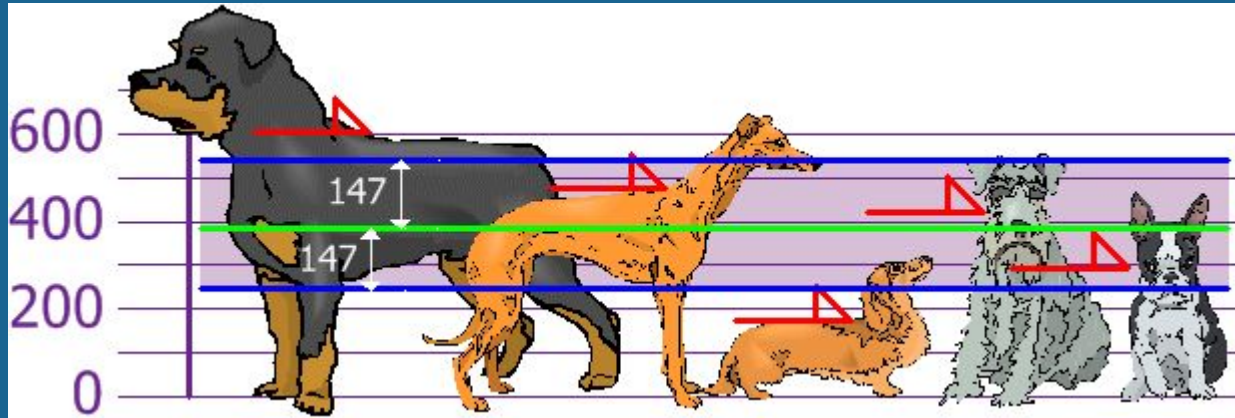


To calculate the Variance, take each difference, square it, and then average the result:

$$\begin{aligned} \text{Variance} &= \text{square}(\sigma) \\ &= 206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2 / 5 \\ &= 108520 / 5 \\ &= 21704 \end{aligned}$$

$$\sigma = \sqrt{21704} = 147$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers are tall dogs. And Dachshunds are a bit short, right?

We can expect about 68% of values to be within plus-or-minus 1 standard deviation.

But ... there is a small change with **Sample Data**

Our example has been for a **Population** (the 5 dogs are the only dogs we are interested in).

But if the data is a **Sample** (a selection taken from a bigger Population), then the calculation changes!

When you have "N" data values that are:

- The **Population**: divide by **N** when calculating Variance (like we did)
- A **Sample**: divide by **N-1** when calculating Variance

Example: if our 5 dogs are just a sample of a bigger population of dogs, we divide by 4 instead of 5, hence:

Sample variance = $108520/4 = 27130$

Sample Standard Deviation = Sq root of 27,130 = 165 (nearest value)

The "**Population** Standard Deviation": $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$

The "**Sample** Standard Deviation": $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$

Looks complicated, but the important change is to divide by **N-1** (instead of **N**) when calculating a Sample Variance.

Why $n-1$ & why not n ??

This is actually called Bessel's correction. The idea behind this is that this is a more unbiased measure of variance than the usual definition.

Imagine you have a huge bookshelf. You measure the total thickness of the first 6 books and it turns out to be 158mm. This means that the mean thickness of a book based on first 6 samples is 26.3mm.

Now you take out and measure the first book's thickness (one degree of freedom) and find that it is 22mm. This means that the remaining 5 books must have a total thickness of 136mm

Now you measure the second book (second degree of freedom) and find it to be 28mm. So you know that the remaining 4 books should have a total thickness of 108mm .

.

.

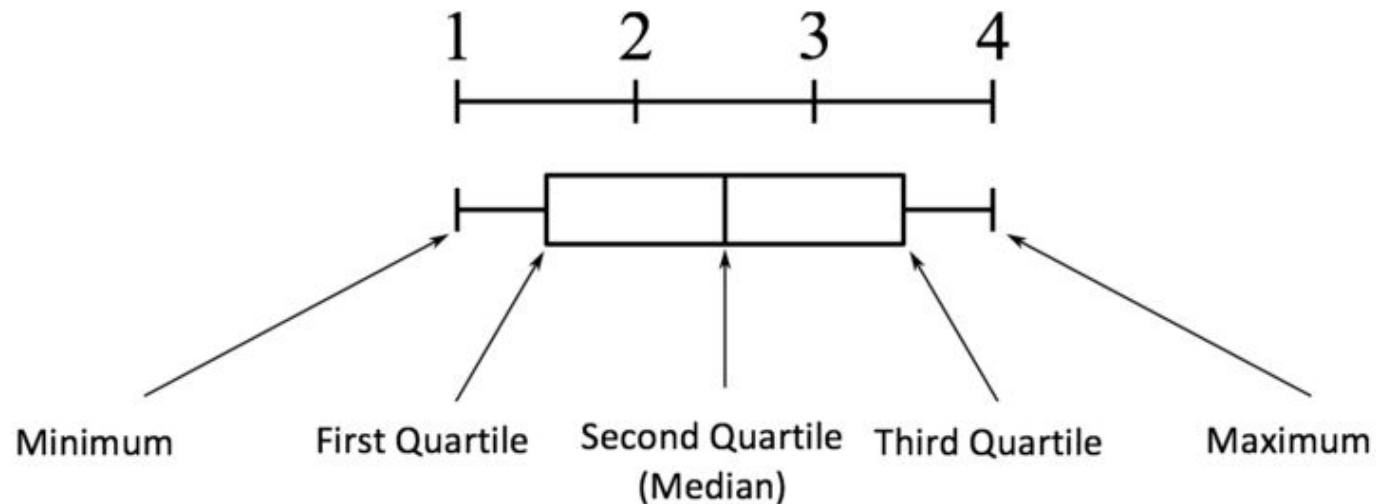
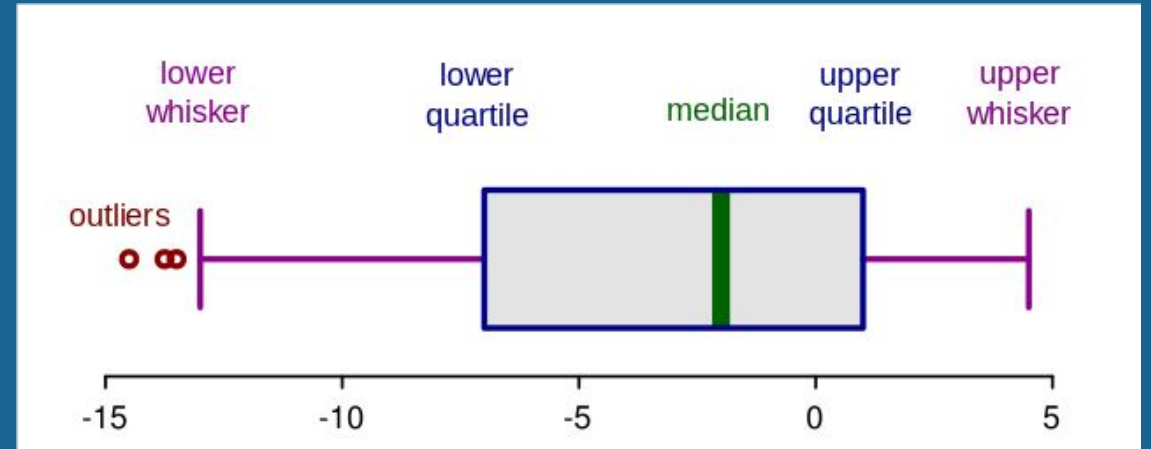
In this way, by the time you measure the thickness of the 5th book individually (5th degree of freedom) , you automatically know the thickness of the remaining 1 book.

This means that you automatically know the thickness of 6th book even though you have measured only 5. Extrapolating this concept, In a sample of size n , you know the value of the n 'th observation even though you have only taken $(n-1)$ measurements. i.e, the opportunity to vary has been taken away for the n 'th observation.

This means that if you have measured $(n-1)$ objects then the n th object has no freedom to vary. Therefore, degree of freedom is only $(n-1)$ and not n .

Detecting Outliers:

2. Boxplots: Box plots are a graphical depiction of numerical data through their quantiles.



Detecting Outliers:

3. Clustering Techniques: There are various clustering techniques such as KMeans, Hierarchical & DBScan, and all of them can be used to identify outliers

4. ML Algorithms

Measure of Central Tendency

Central tendency measures the center value of the dataset .It give us idea about the concentration of the value in the central part of the distribution.



Mean/Average

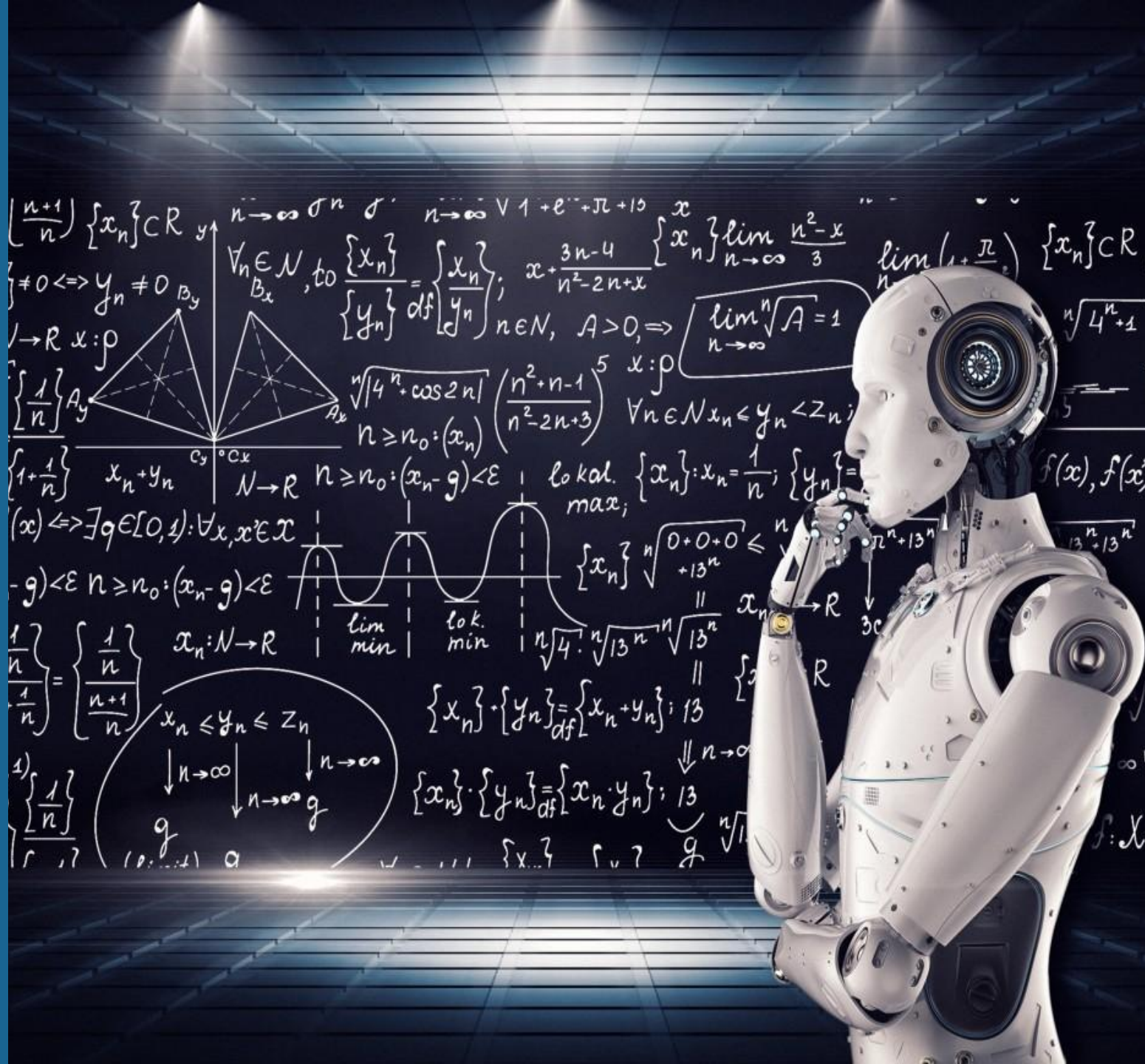
It's an average set of observation of the data. It compute the sum of all observation present in the datasets divided by total number of observation.

Steps to find the Mean/Average:

1. Add/Sum each number/observation present in the dataset.
2. Calculate the total number present in the dataset.
3. Divide sum of observation to the total number of observation.

Formula:-

$$\bar{X} = \frac{\sum X}{N}$$



Median

Median is the middle value of the entire dataset. It splits the whole dataset into two parts and takes the middle value of the datasets. It's also called the 50th percentile.

Steps to find Median of the datasets:-

(For odd)

1. First Arrange the observation in Increasing order.
2. Divide the observation into two equal parts 50/50.
3. Take the middle value of the data.

(For even)

1. First arrange the values in increasing order.
2. Divide the observation into two equal parts 50/50.
3. Now take the middle two values of the dataset which remain after dividing.
4. Now take the average of the middle two values, that is the median of the dataset.

In case of discrete distribution the formula is:

It is obtained by considering the Cumulative frequency

$$N/2$$

In case of continuous distribution:

$$\text{Median} = l + \frac{h}{f} \left(\frac{N}{2} - c \right)$$

Where:

l = lower class boundary of the median class

h = Size of the median class interval

f = Frequency corresponding to the median class

N = Total number of observations i.e. sum of the frequencies

c = Cumulative frequency preceding median class.

Mode

Mode is the value which occur most frequently in the set of the observation. Data can have more than one mode as Uni-model, Bi-model, Multi model. It's Usage depends on the on situation as Max(),Min(),Mean().

Steps for finding the Mode:-

It's very easy to find mode of any observation

1. Take the Most frequent value present in the dataset.

Special Cases:

1. If the maximum number of frequency repeated
2. If the maximum frequency is occur at the beginning and end of the observation.
3. If there is irregularities in the distribution.

In all the above cases we find the mode of the observation by using method of grouping.

In case of continuous distribution the formula is:

Measures of Dispersion

Measure of dispersion indicates that how the data is dispersed from the measure of central tendency.

01

Range



02

Inter-Quartile range



03

Variance



Standard deviation



05

Mean Deviation



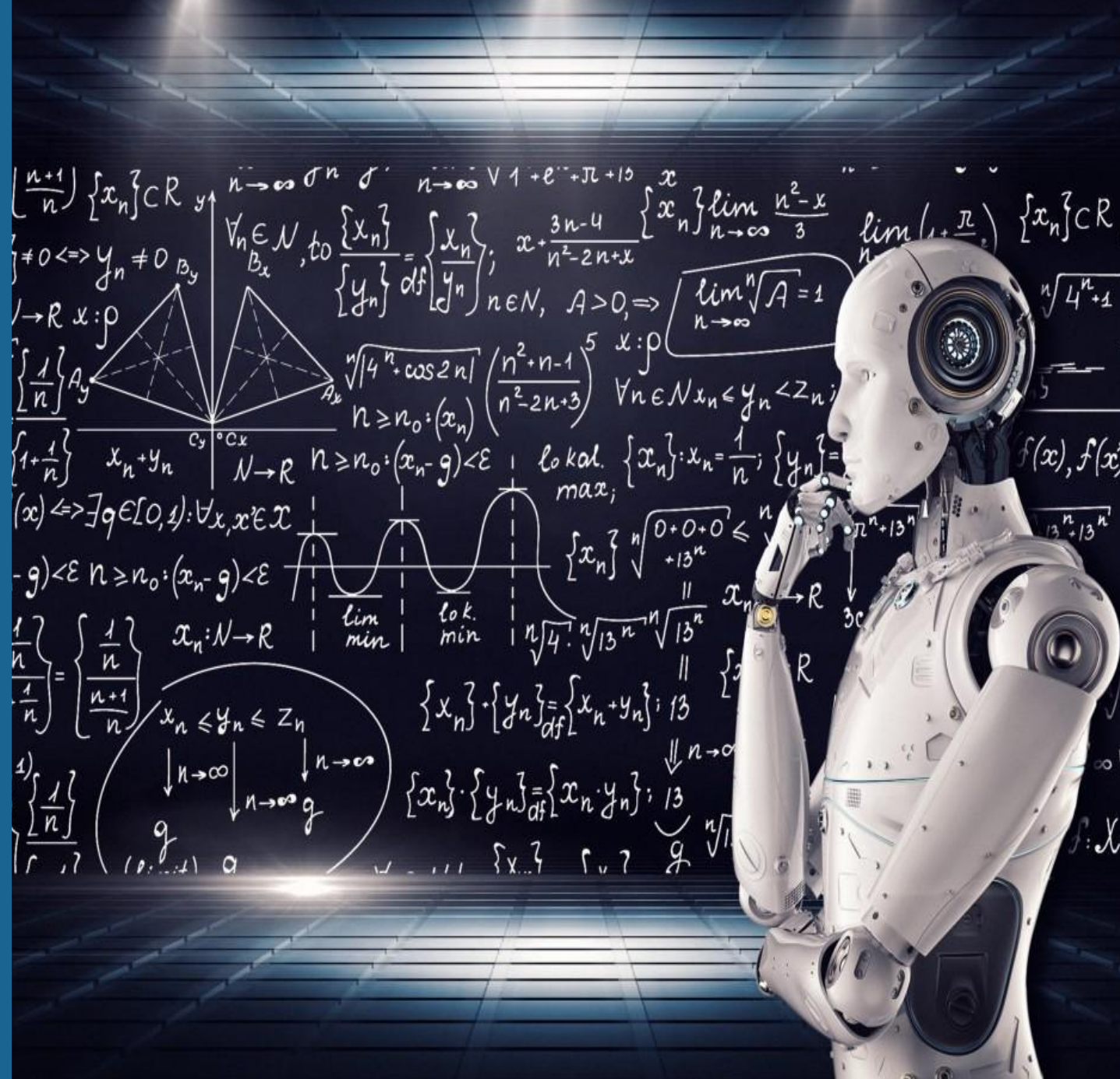
Range

Range is simplest measure of dispersion. It's measure the difference between highest value and lowest value present in the dataset.
It's used to construct control chart in quality assurance.

The formula of Range is:

Range = Highest value – lowest value

Range is useful when you want to focus on extremes values of the dataset.

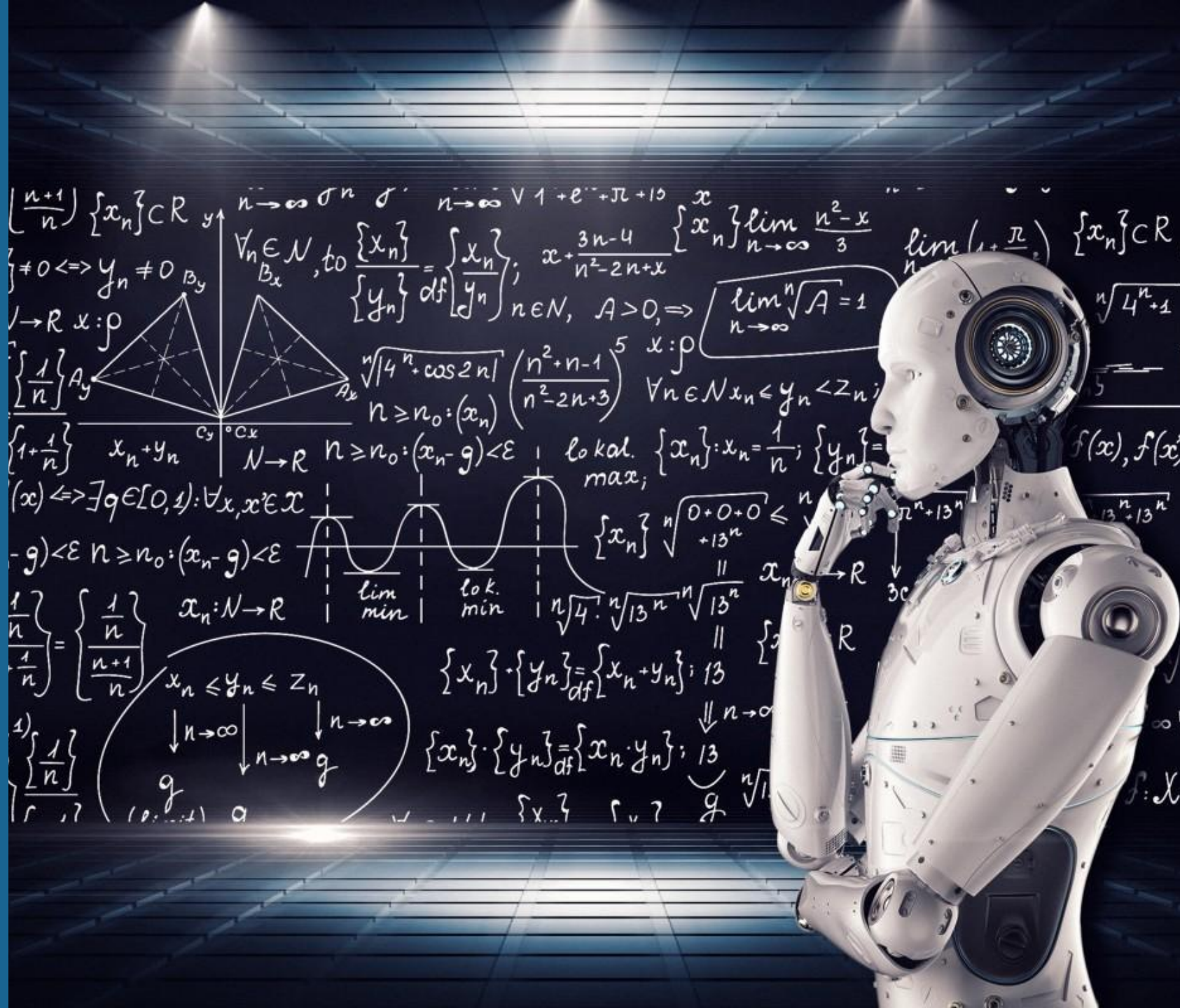


Inter-Quartile Range

- Inter-quartile range measure the middle 50% of the data.
- Inter-quartile range indicate how the data is dispersed around the mean.
- It's a difference between the third quartile and first quartile value of the dataset.
- IQR is helpful to detect the outlier present in the dataset.

The Formula of IQR is:

$$\text{IQR} = Q_3 - Q_1$$



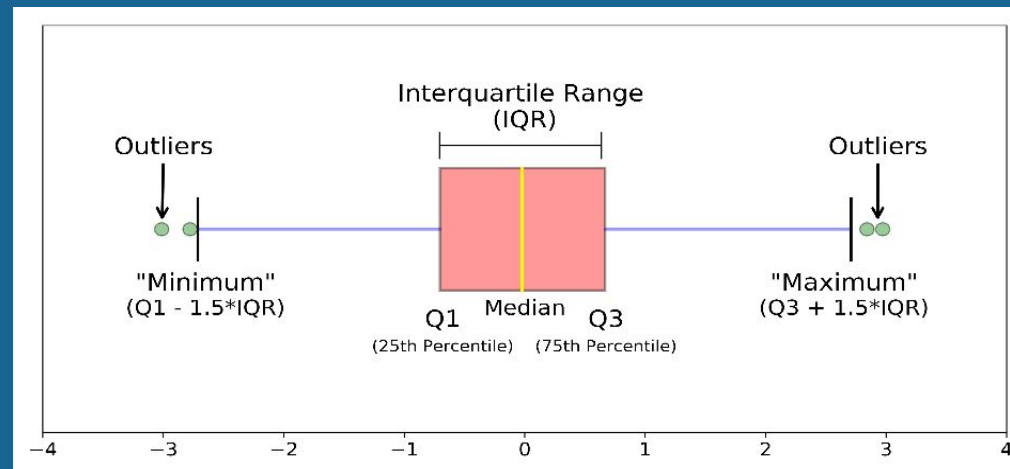
Example

Suppose we have a data series 88,89,89,89,90,91,91,91,92

So, to find out the IQR first we have to sort the data on ascending order as the data is already sorted so we don't need to sort it. Now next find the median (middle value) of the data this is identified as **Q2**, the middle value of the dataset is 90.

As the dataset is divided into two parts, now find the middle value of the first half which is identified as **Q1** is 89 and second half which is identified as **Q3** is 91.

So, the IQR is $Q_3 - Q_1 = 91 - 89 = 2$.



- Variance measure the dispersion of the data around the mean of the data.
- Variance indicate how the data is dispersed from its mean.
- If the value of variance is closer to mean then it's a low variance.
- If there is significant difference in the value from the mean then it's a high variance.
- Variance is denoted by σ^2 (sigma square).

Formula for population variance:

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

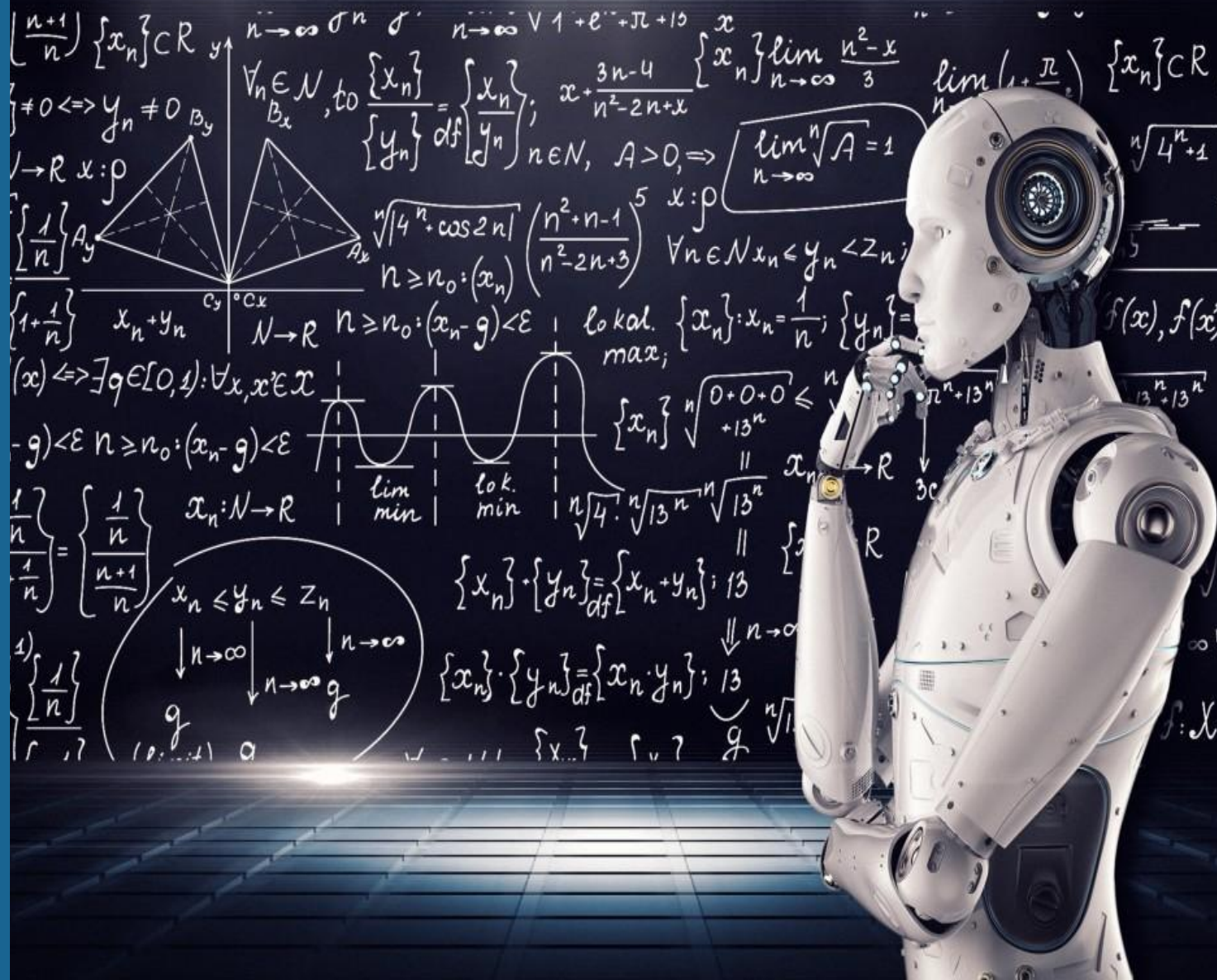
where N is the population size and the X are data points and μ is the population mean.

Formula for sample variance:

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

where n is the sample size and X are the data points and \bar{x} is the sample mean.

Variance



- Standard deviation is most important and frequently method in measure of dispersion.
- Standard deviation is simply the Square root of the variance.
- Standard deviation indicate how far away the datapoints is dispersed from the mean.
- It is denoted by σ .

The formula of standard deviation for population:

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

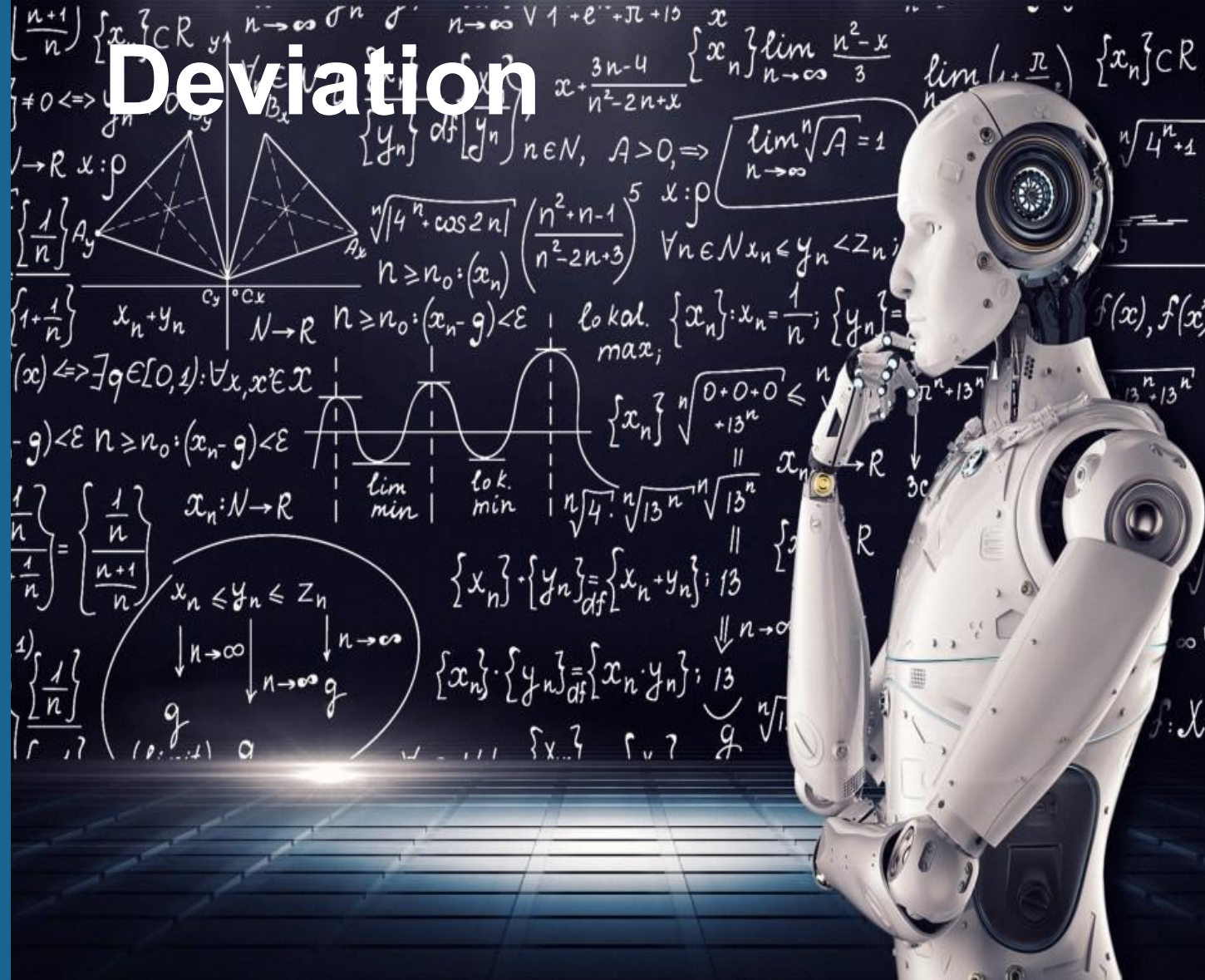
where **N** is the population size and the **X** are data points and μ is the population mean.

The formula of standard deviation for sample:

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

where **n** is the sample size and **X** are the data points and \bar{x} is the sample mean.


Standard Deviation



Example

Suppose I am travelling from Indore to Bhopal by car, my car speed data is 4,6,3,5,2 the average speed of car is **4**. Now we calculate **variance** of car speed data, we get the variance **2**(by population formula).

Standard deviation :the variance is 2, we calculate the standard deviation which is **1.41**, so this indicates that our data is fluctuate in between **4 ± 1.41** (if take one standard deviation, that is 68% of the total data).

Values	Mean – Value	Square
4	0	0
6	-2	4
3	1	1
5	-1	1
2	2	4
Mean = 4		Sum=10
Variance = 2		
	Std. Dev = 1.41	

Mean Deviation

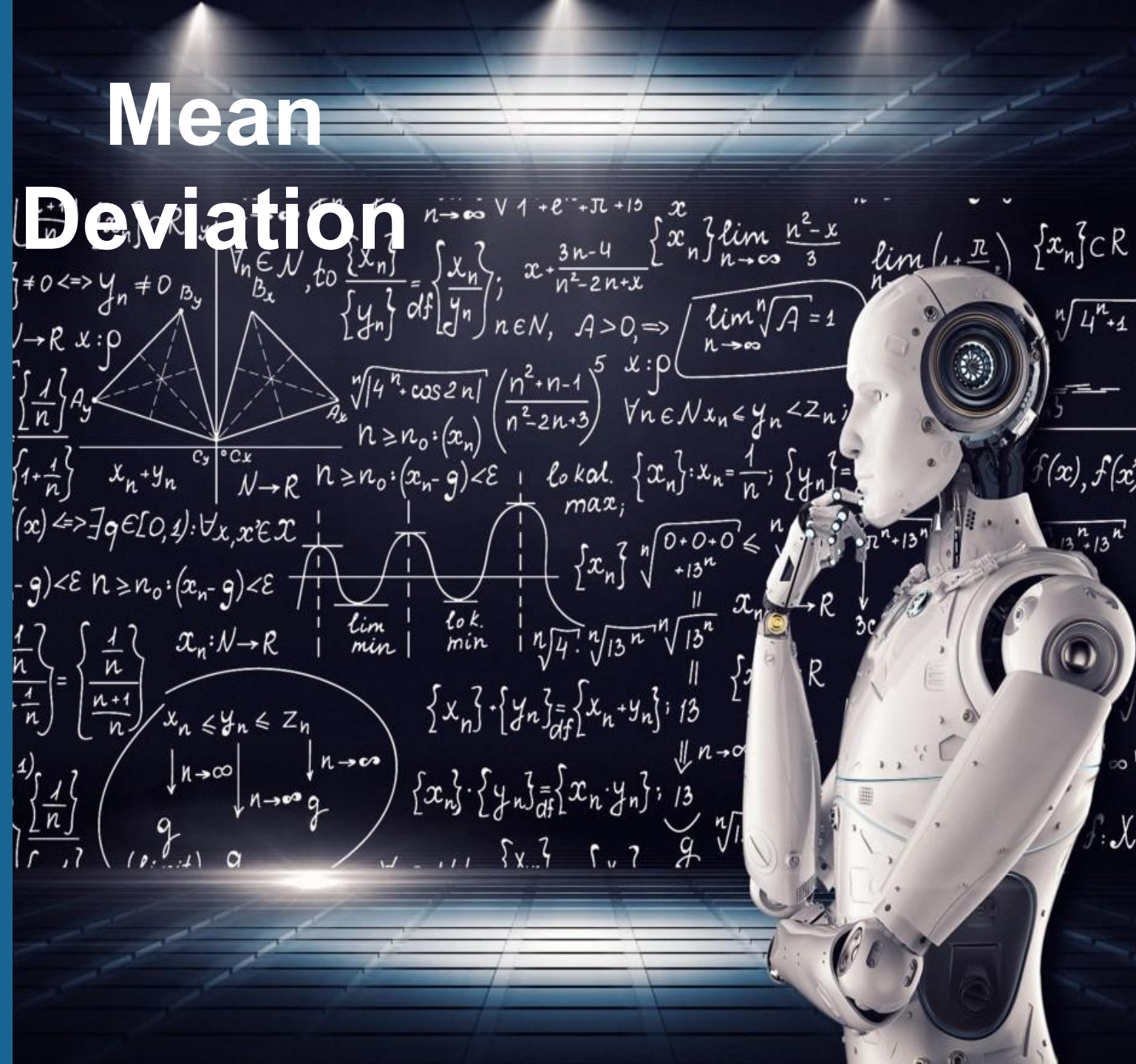
Mean deviation is defined as the average sum of the absolute values of the deviation from any arbitrary values viz. mean, median, mode etc. Its often suggested it to calculate from the median because its give least value when measured from the median.

The deviation of an observation x_i from the assumed mean A is defined as

$$(x_i - A)$$

Therefore the mean deviation can be defined as

$$\sum(|x_i - A|)/n$$



Skewness

Skewness measure the distribution of the data. It indicate weather the data is distributed symmetric or not. If the data is distributed symmetric means the data is normally distributed.

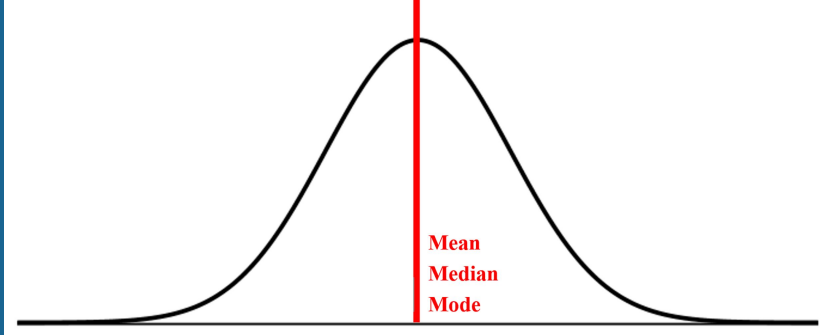
Types of Skewness:

1. Symmetric
2. Positive skewness
3. Negative skewness

Pearson coefficient of skewness:-

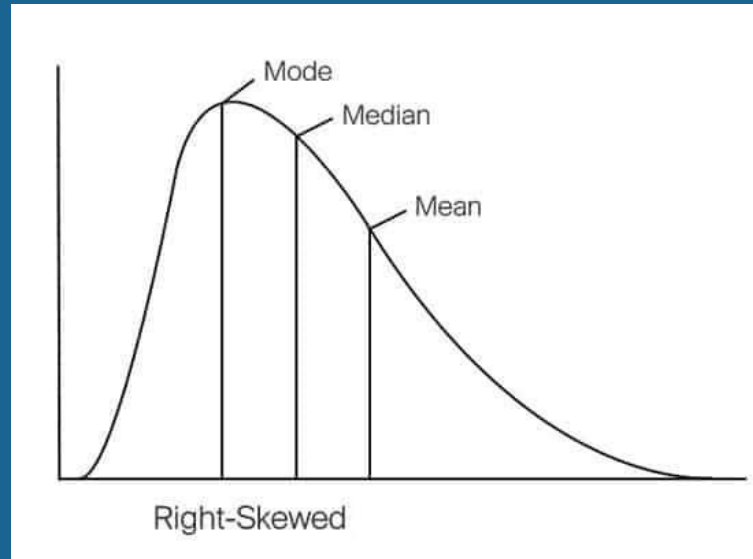
$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{standard deviation}}$$





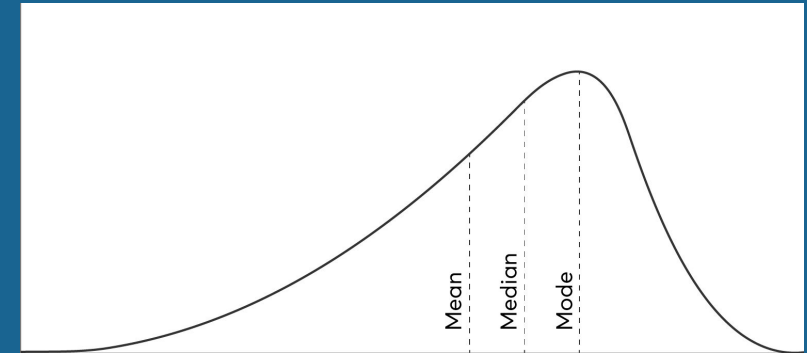
Normally distributed (Symmetric)

Mean = Median= Mode



Positive skewed(right skewed)

Mean>Median>Mode



Negative skewed (Left skewed)

Mean<Median<Mode

Kurtosis

Kurtosis measure the thickness and flatness of the distribution of the data. The degree of tailedness of the data is measure by the kurtosis. It tells us the extent to which the distribution is more or less prone than the normal distribution.

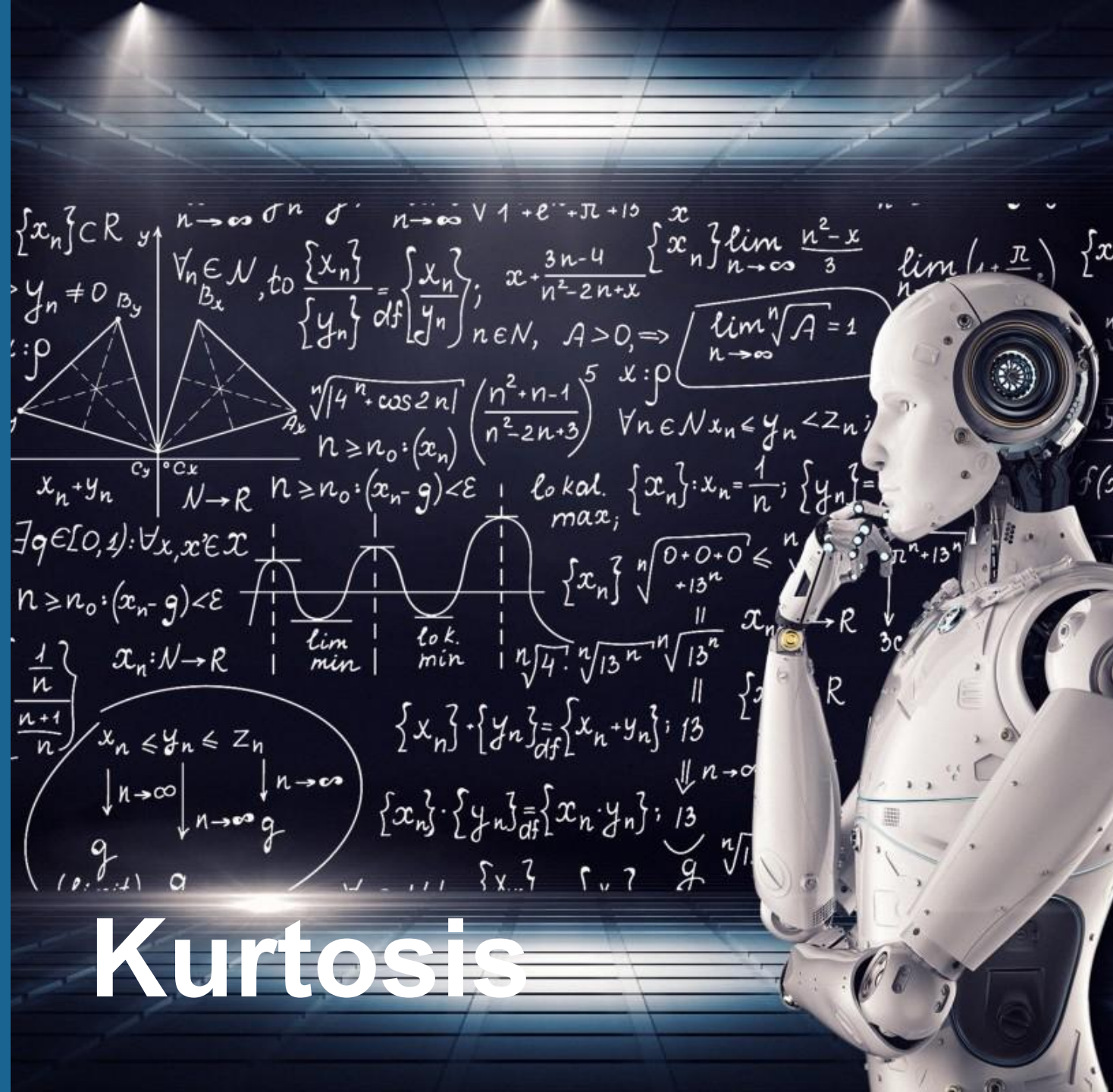
Type of Kurtosis:-

1. Platykurtic
2. Mesokurtic
3. Leptokurtic

It is Measured by Coefficient β_2 .

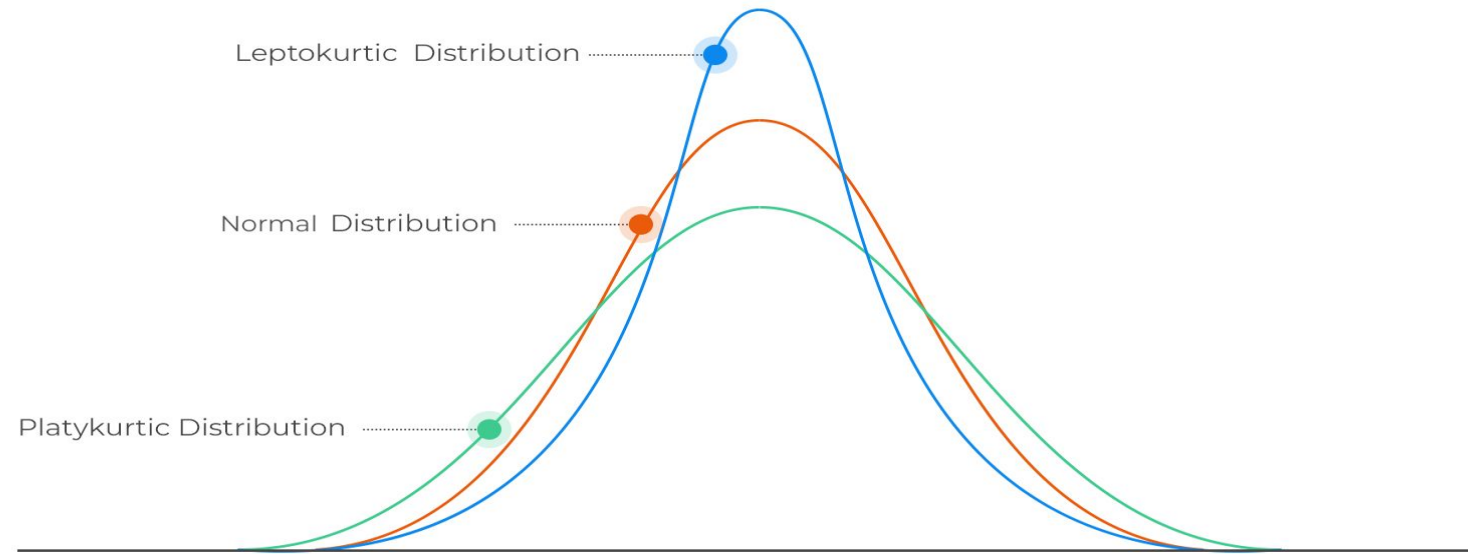
$$\beta_{2.} = \mu_4 / \mu_2^2$$

$$\gamma_2 = \beta_2 - 3$$



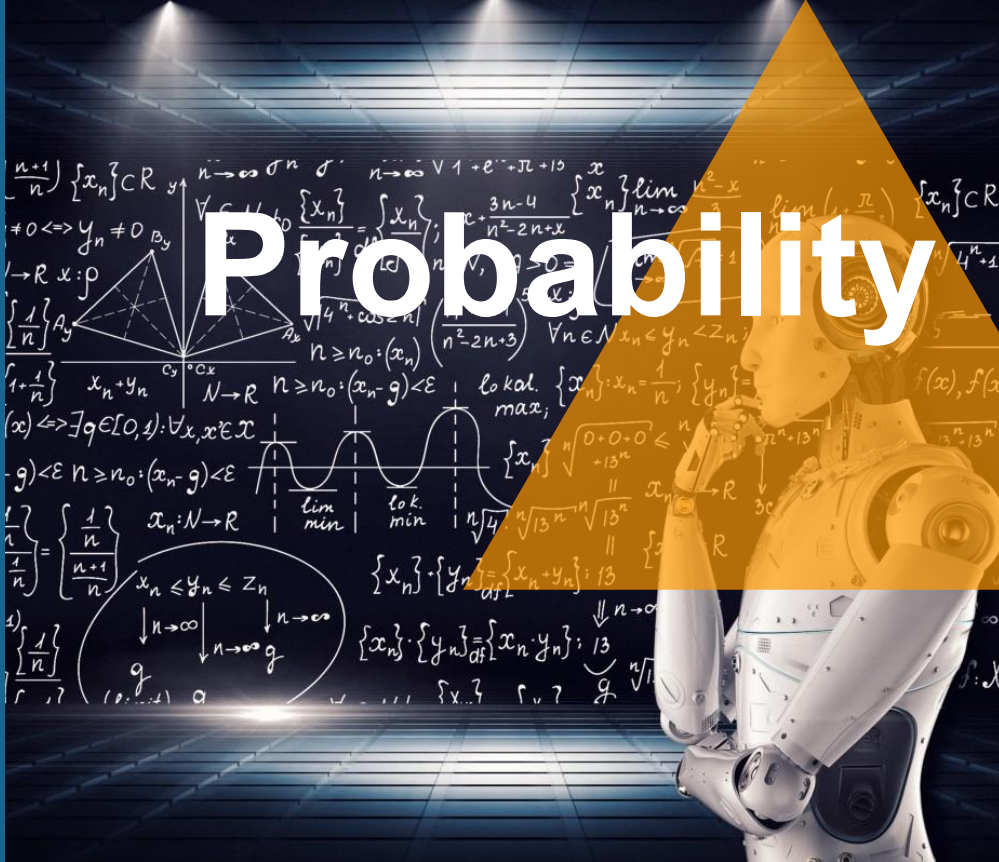


Kurtosis



Kurtosis: 0, means Normal distribution
Kurtosis: Negative, Platykurtic distribution
Kurtosis: Positive, Leptokurtic distribution

Probability



If the trial is repeated number of times under homogenous and identical condition then the value of ratio of number of favourable cases to the total number of possible cases is called probability.

$$\text{Probability} = \frac{\text{No. of Favourable outcome}}{\text{Total no. of possible outcome}}$$

Probability is always in between 0 to 1, which measure how likely an event to be occur.

Trial ➤ Each performance in a random experiment

Event ➤ Outcome of trial is an event $P(E)$.

Random Experiment ➤ To perform more than once

Sample Space ➤ Set of all possible outcome in a random experiment $P(s)$.

Example

Question 1 :What is the Probability of getting an even when throwing a single die.

Solution- Sample Space (S) = {1,2,3,4,5,6}

Event (A) = getting an even number

$P(A) = ?$

$A = \{2,4,6\}$

$$P(A) = \frac{\text{No.of favourable cases}}{\text{Total no of possible outcome}}$$

$$P(A) = 3/6$$

$$P(A) = \frac{1}{2}$$

So probability of getting a even number in single die is $\frac{1}{2}$.

Question 2 :What is the Probability of getting an head when toss a fair coin.

Solution- Sample Space (S) = {H,T}

Event (A) = getting an Head

$P(A) = ?$

$A = \{H\}$

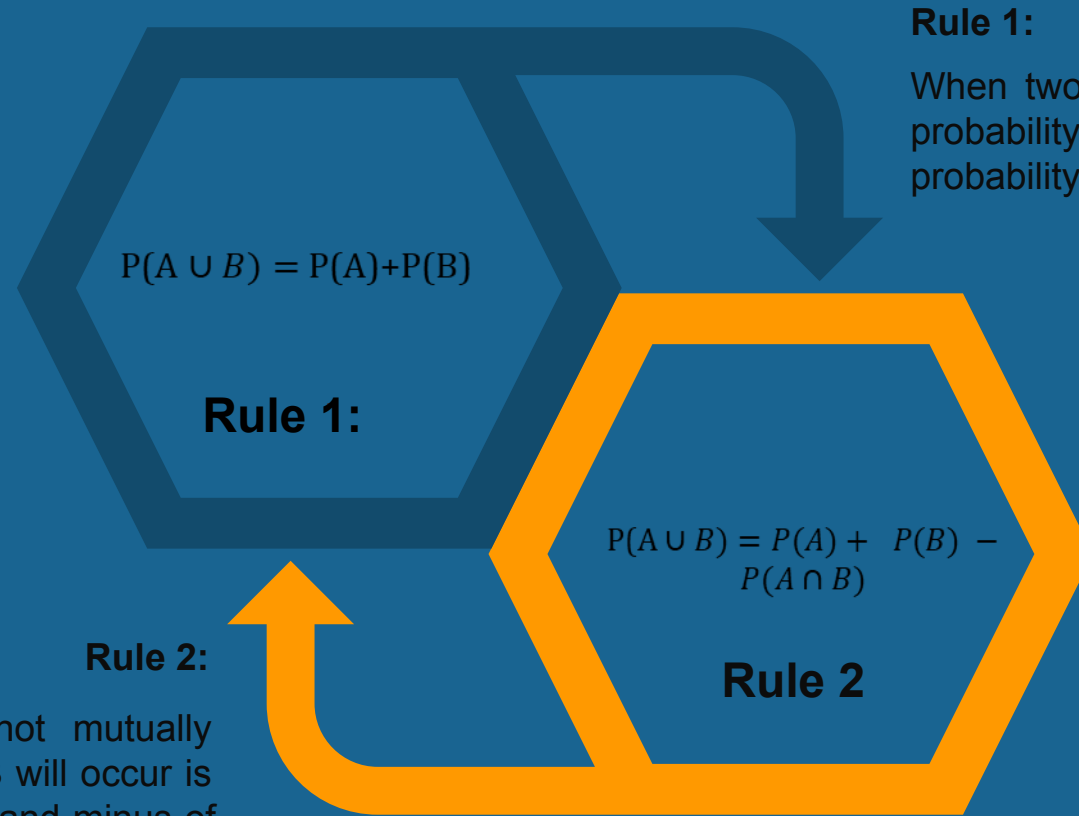
$$P(A) = \frac{\text{No.of favourable cases}}{\text{Total no of possible outcome}}$$

$$P(A) = 1/2$$

$$P(A) = \frac{1}{2}$$

So probability of getting a head on tossing a fair coin is $\frac{1}{2}$.

Addition Rule of Probability



Rule 1:

When two events A & B are mutually exclusive, the probability that A or B will occur is the sum of probability of each event..

Rule 2:

When two events A & B are not mutually exclusive, the probability that A or B will occur is the sum of probability of each event and minus of intersection of A and B.

Mutually exclusive= Intersection is empty.

Example

Question 1: A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?

Solution: Sample Space (S) = {1,2,3,4,5,6}

Event (A) = getting an even number

P(A) = ?, P(B) = ?

$$P(A \cup B) = P(A) + P(B)$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{1}{3}$$

Question 2: In a math class of 30 students, 17 are boys and 13 are girls. On a unit test, 4 boys and 5 girls made an A grade. If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?

Solution: $P(A \cap B) = P(A) + P(B) - P(A \cap B)$

$$P(\text{girl}) = 13/30,$$

$$P(A) = 9/30$$

$$P(A \cap B) = 5/30$$

$$P(\text{girl or A}) = P(\text{girl}) + P(A) - P(\text{girl and A})$$

$$= 13/30 + 9/30 - 5/30$$

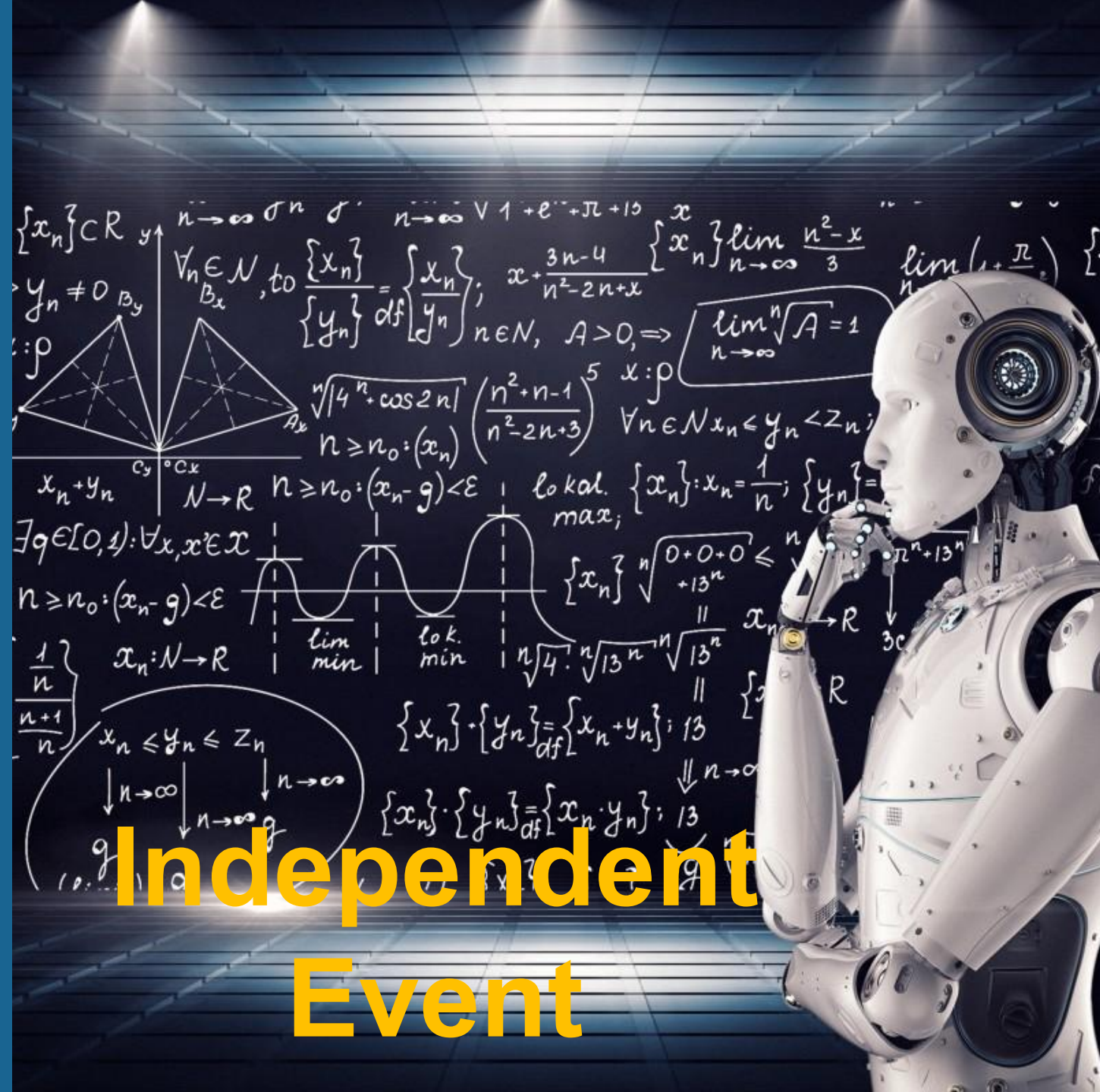
$$= 17/30$$

Independent Events:

When two events A and B are said to be independent if any of the following equivalent statement hold:

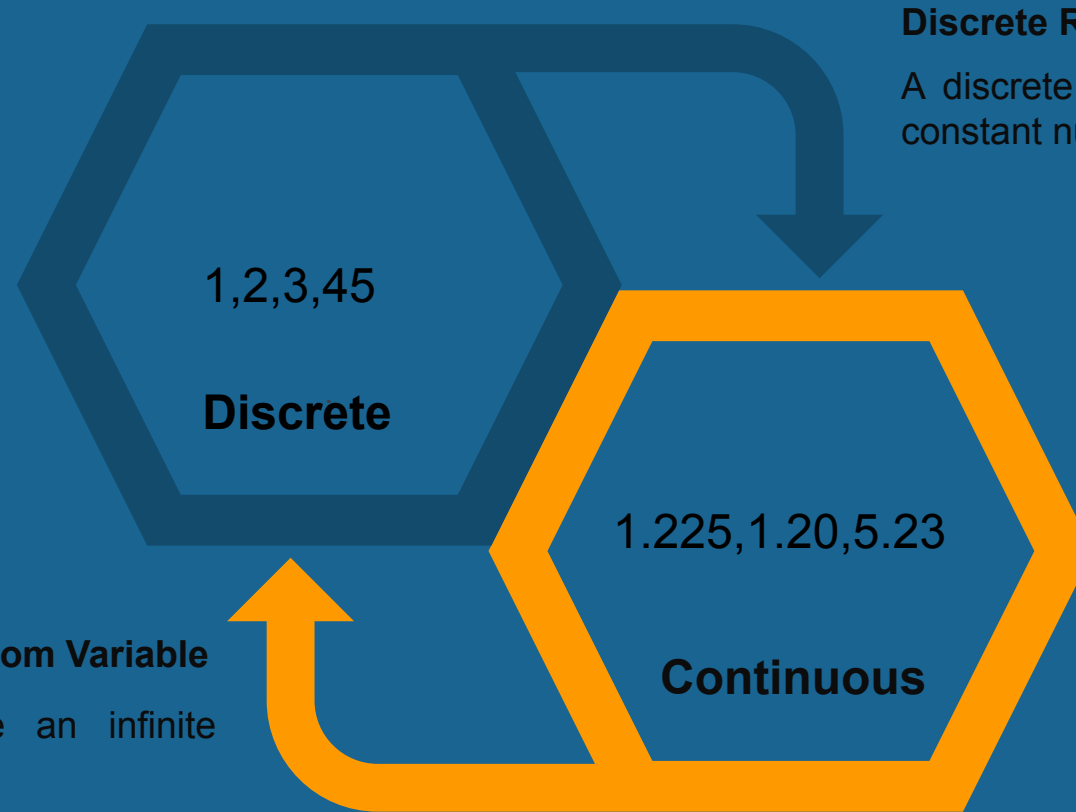
- $P\left(\frac{A}{B}\right) = P(A)$
- $P\left(\frac{B}{A}\right) = P(B)$
- $P(A \& B) = P(A) \cdot P(B)$ (Multiplication Rule for independent.)

Two events, A and B, are **independent** if the fact that A occurs does not affect the probability of B occurring.



Random Variable

A random variable is a variable which is associated with the outcome of random experiment.



Discrete Random Variable

A discrete random variable is one which take only constant number of value.

Continuous Random Variable

Continuous random variable have an infinite continuum number of possible value.

Mutually exclusive= Intersection is empty.

Cumulative Probability

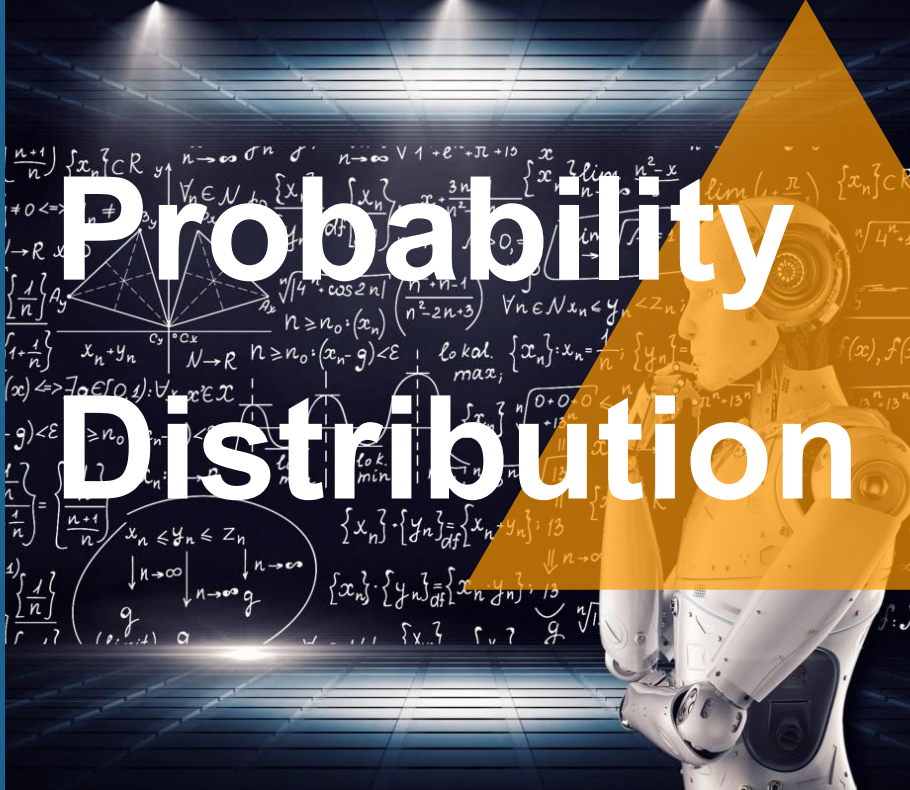


Cumulative probability refers to the probability that the value of a random variable falls within a specified range. In other word its defined as the probability of variable being less than or equal to specified value.



Example: Consider a coin flip experiment. If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads? The answer would be a cumulative probability. It would be the probability that the coin flip results in zero heads plus the probability that the coin flip results in one head. Thus, the cumulative probability would equal:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$$



The distribution of the probability across the range of possible value is known as probability distribution of random variable.

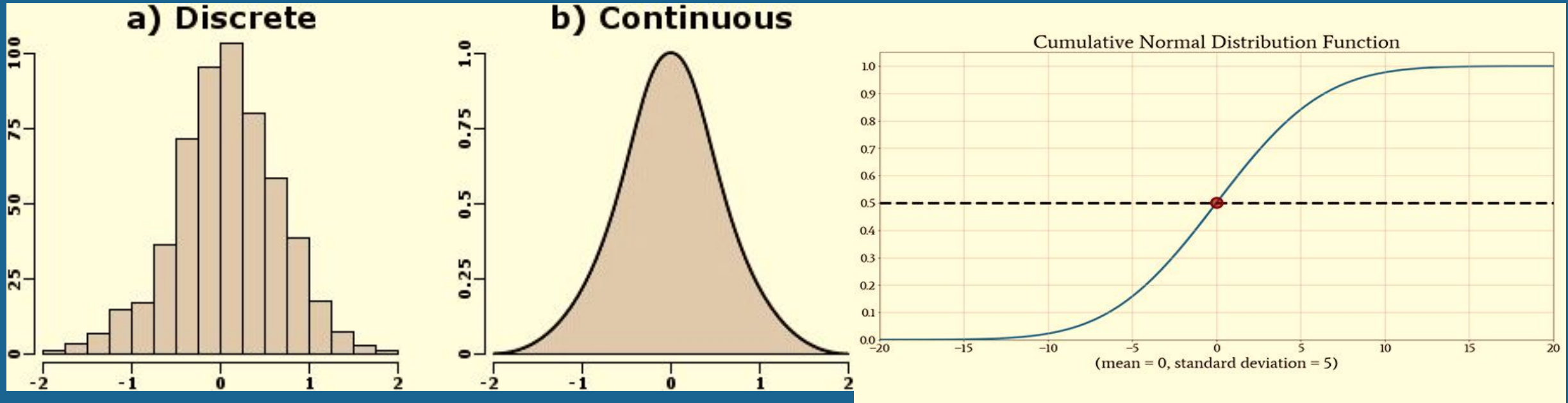
Probability distribution is ANY form of representation that tells us the probability for all possible values of X. It could be any of the following:

- A table
- A chart
- An equation- $P(x) = x/21$ (for $x = 1, 2, 3, 4, 5$ and 6)

Continuous Distribution ➤ Probability distribution of continuous random variable are called continuous distribution or **Probability density function(PDF)**.

Discrete Distribution ➤ Probability distribution of discrete random variable are called discrete probability distribution or **probability mass function(PMF)**.

Cumulative Distribution Function ➤ Cumulative distribution function is distribution which plot cumulative probability of X against X.



Age, is it continuous or discrete

Example: Anybody can be having 22.67 years age, or 22.6789 years age, if we consider days, years, weeks etc, in that case its a continuous feature.

But normally we consider everyone with 1-5 years age group to be under 5, and so on, so in this case, it can be considered as discrete

- **N** identical trials are performed where **n** is determined prior to the experiment.
- Trials are independent
- Each trial has two possible outcomes which we call success or failure.
- The probability of success is the same for each trial and is denoted by “p”. Thus for each trial

P(Success) = p

P(failure) = q = 1-p

- Random variable is:
x = the total number of success in “n” trial.
- $X \sim \text{Bin}(n, p)$

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where

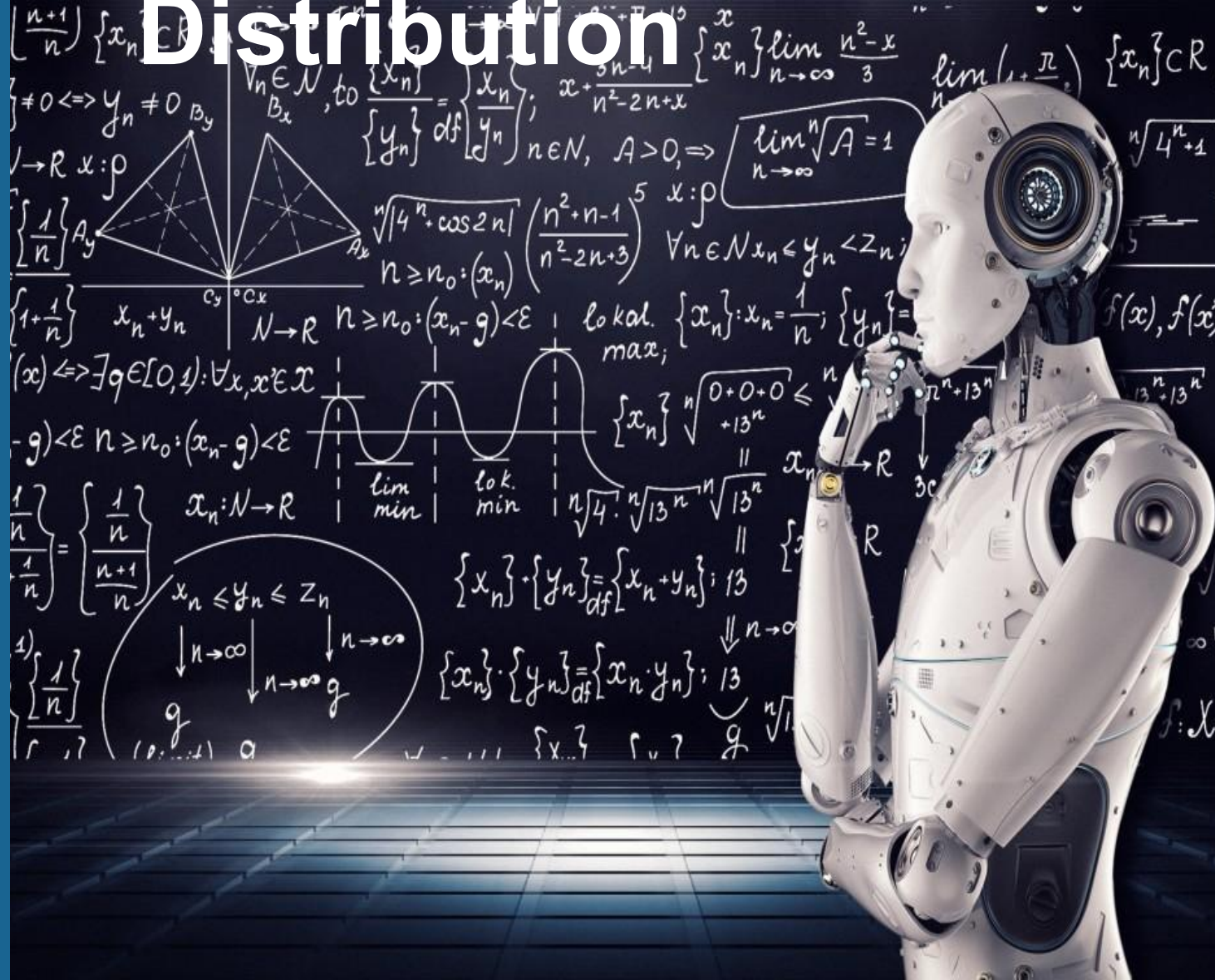
n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

q = 1 - p = the probability of getting a failure in one trial

Binomial Distribution



Example

Question: A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

No. of trial (n) = 10

X = 6

P(success) = probability getting a 6 head = p = 0.5

P(failure) = probability of not getting a head = q = 0.5

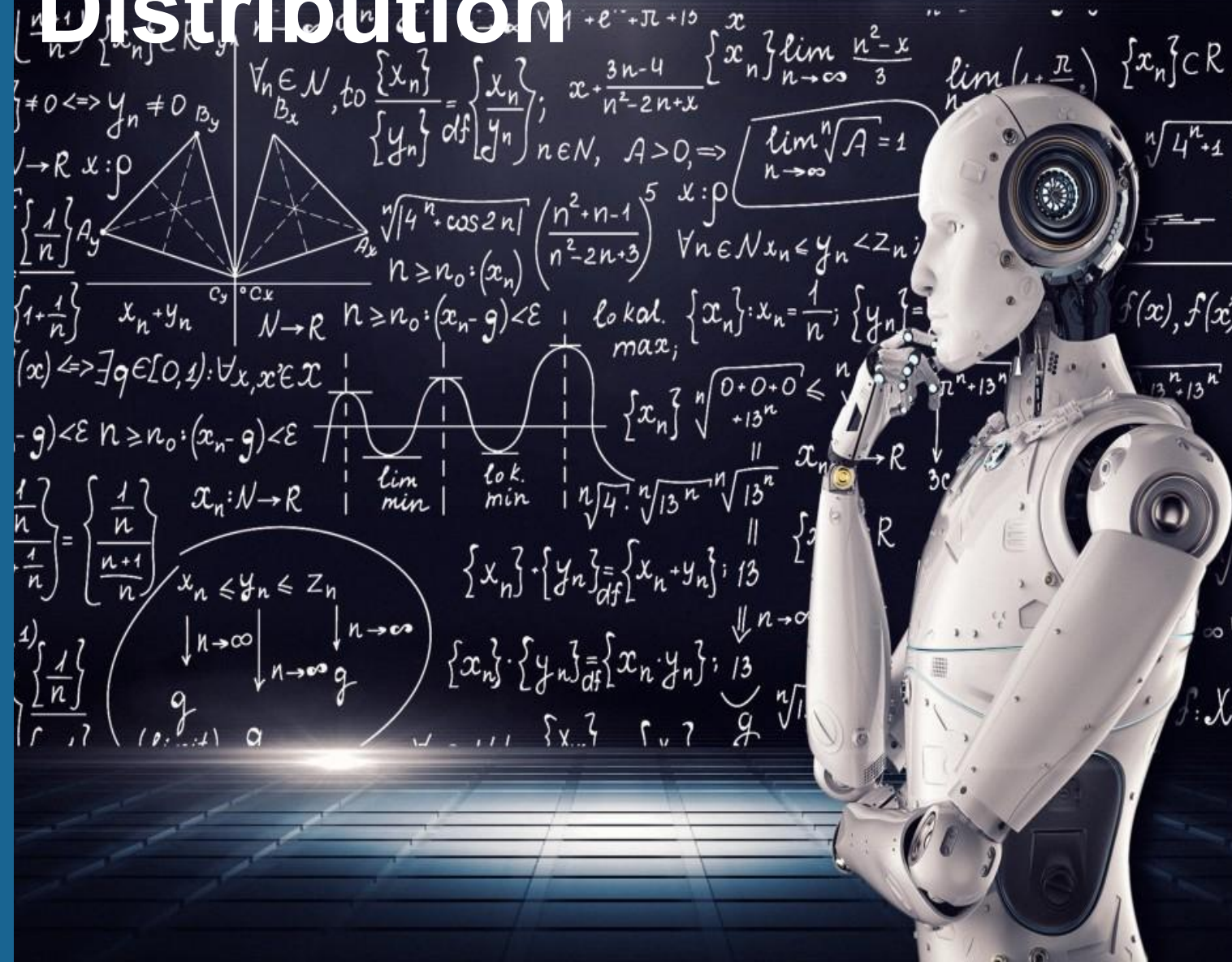
By binomial distribution,

$$P(x) = \frac{10!}{(4! * 6!)} * 0.5^6 * 0.5^4 \\ = 0.2$$

$$P(x) = {}_n C_x p^x (1 - p)^{n-x}$$

The Poisson parameter Lambda (λ) is the total number of events (k) divided by the number of units (n) in the data ($\lambda = k/n$).

Poisson Distribution



Example

Question: Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district .What is the probability that exactly 5 houses will have fire during the year. Given that $e^{-2} = 0.13534$

Solution: - $n = 2000$, $p = 1/1000$,
 $\lambda = np$
 $= 2000 * 1/1000$
 $= 2$

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(X=5) = \frac{2^5 e^{-2}}{5!}$$

$$= 0.0360$$

probability that exactly 5 houses will have fire during the year is 0.0360.

Type of distribution in which all outcomes are equally likely. Each value has probability that it will be the outcome.

Type of Uniform Distribution:

- **Discrete Uniform:** $X \sim \text{discrete uniform}(a, b)$ is used to indicate that random variable has discrete uniform distribution with parameter a & b where $a < b$. A discrete random variable with parameter a & b has probability mass function.

$$f(x) = 1/(b-a+1), E(x) = (a+b)/2 \\ V(x) = (b-a+1)^2 - 1/12$$

If X is uniformly distributed on the set $\{1, 2, \dots, n\}$ Then:

$$P(X=x) = 1/n, E(x) = (n+1)/2, V(x) = (n^2 - 1)/12$$

- **Continuous Uniform:** It is the probability of random no. selection from continuous interval between a & b . Its density function defined by the following:

$$f(x) = \begin{cases} 1/(b-a), & \text{if } a < x < b \\ 0, & \text{Otherwise} \end{cases}$$

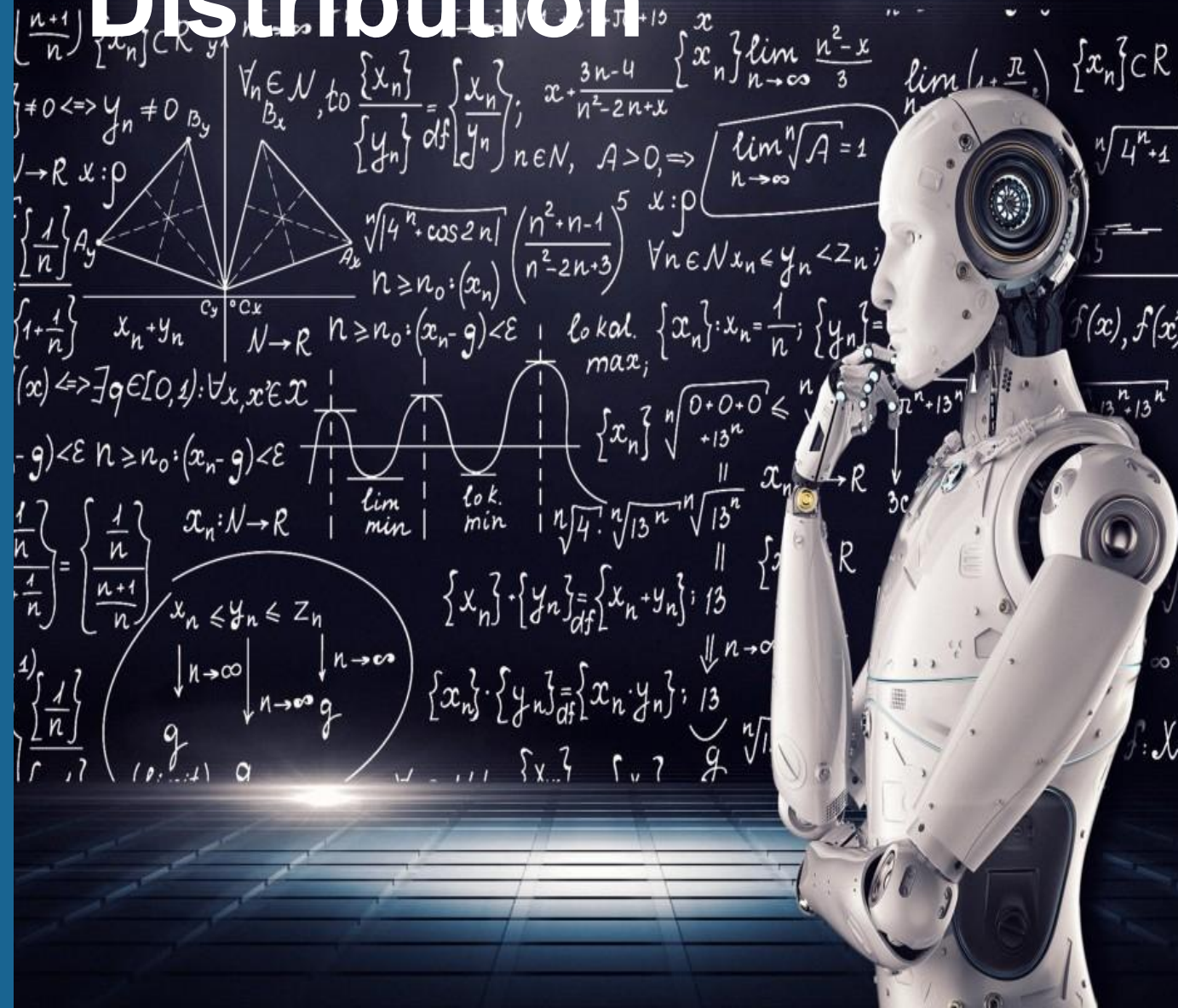
where ,

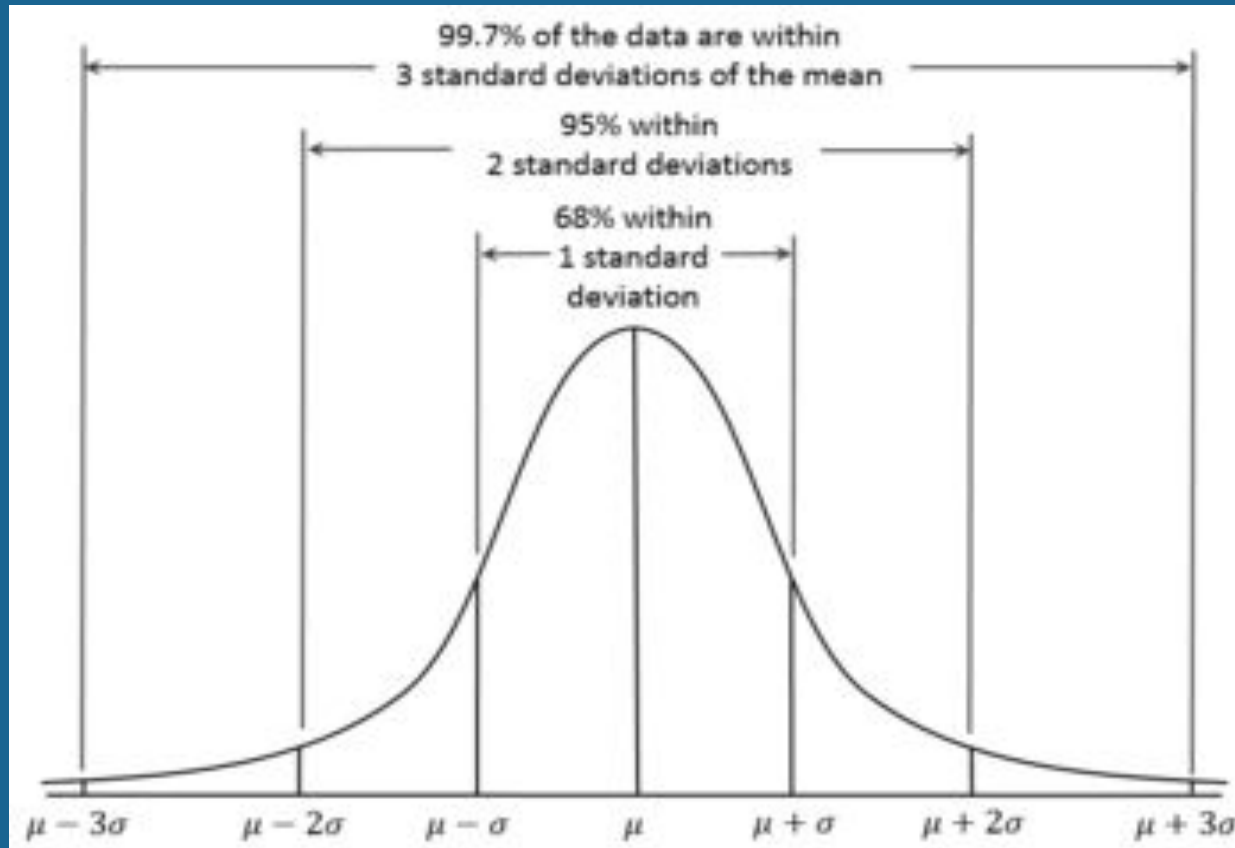
$f(x)$ = value of density function at any X value

a = lower limit of interval

b = upper limit of interval

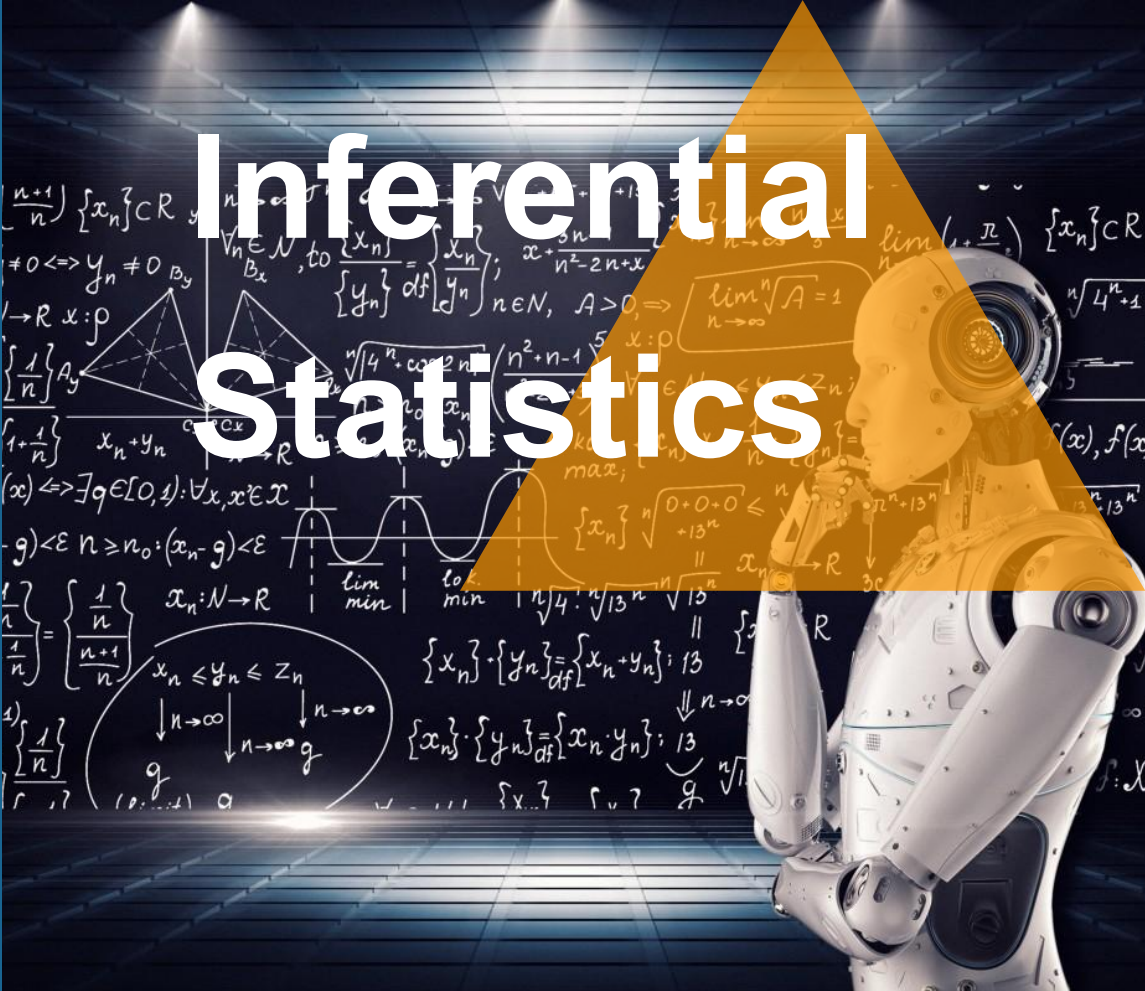
Uniform Distribution





- 68% probability of the variable lying within 1 standard deviation of the mean
- 95% probability of the variable lying within 2 standard deviations of the mean
- 99.7% probability of the variable lying within 3 standard deviations of the mean

Inferential Statistics



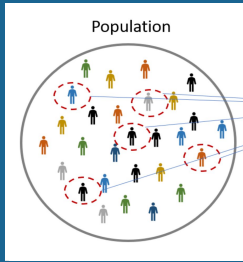
Inferential Statistics

Inferential Statistics is a set of method that is used to draw a conclusion about the characteristics of population based on the sample of the data. It is used to find the population parameter when you have no initial number to start with.

The two main areas of inferential statistics:

1. Estimating parameter
2. Hypothesis Testing

Population And Sample



Population

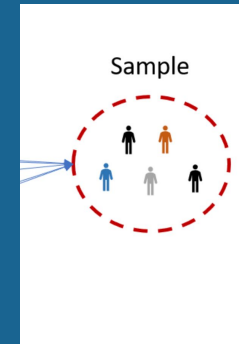
Population

Population is the entire set of respondents. The universe of objects that is required to be analysed is known as population.

Sample

Sample

The subset of population is known as sample. The specific group of individuals that you will collect form the population is known as sample.



Why Sampling is Important?

Gathering data from entire population is not possible. Sampling is applicable in such situation.



Using sampling one can make information faster

Surveying and measuring everyone is not cost effective.



We can easily analyse the data when using sample of the data.

A smaller set of individuals often results in lesser data collection error.



Type of Sampling Technique

```
graph TD; A[Type of Sampling Technique] --> B[Probability sampling]; A --> C[Non-Probability sampling]; B --> D[Random]; B --> E[Systematic]; B --> F[Stratified]; B --> G[Cluster]; C --> H[Convenience]; C --> I[Purposive]; C --> J[Voluntary response]; C --> K[Snowball];
```

Probability sampling

Random

Systematic

Stratified

Cluster

Non-Probability
sampling

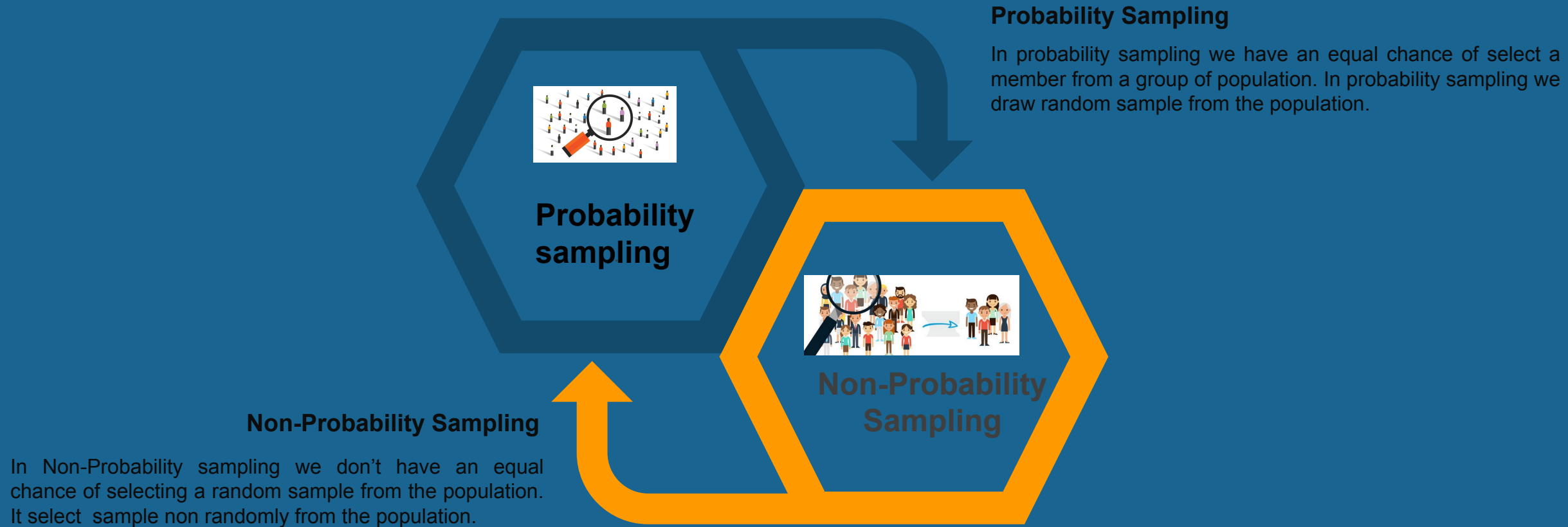
Convenience

Purposive

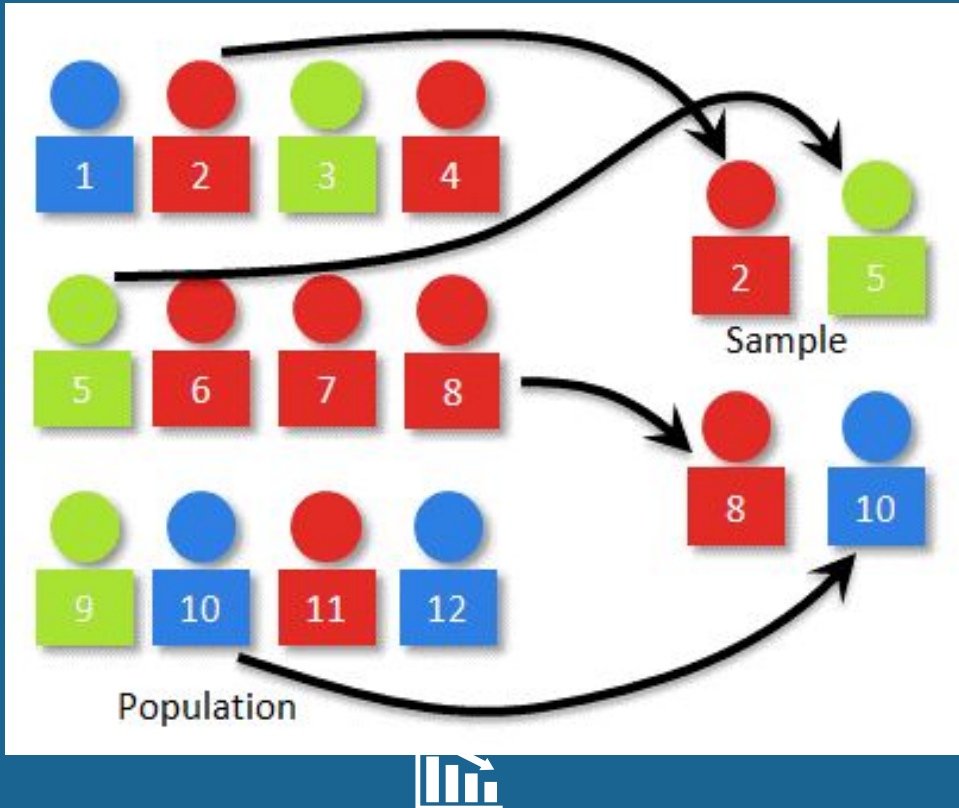
Voluntary
response

Snowball

Probability Sampling and Non-Probability Sampling



Simple Random Sampling

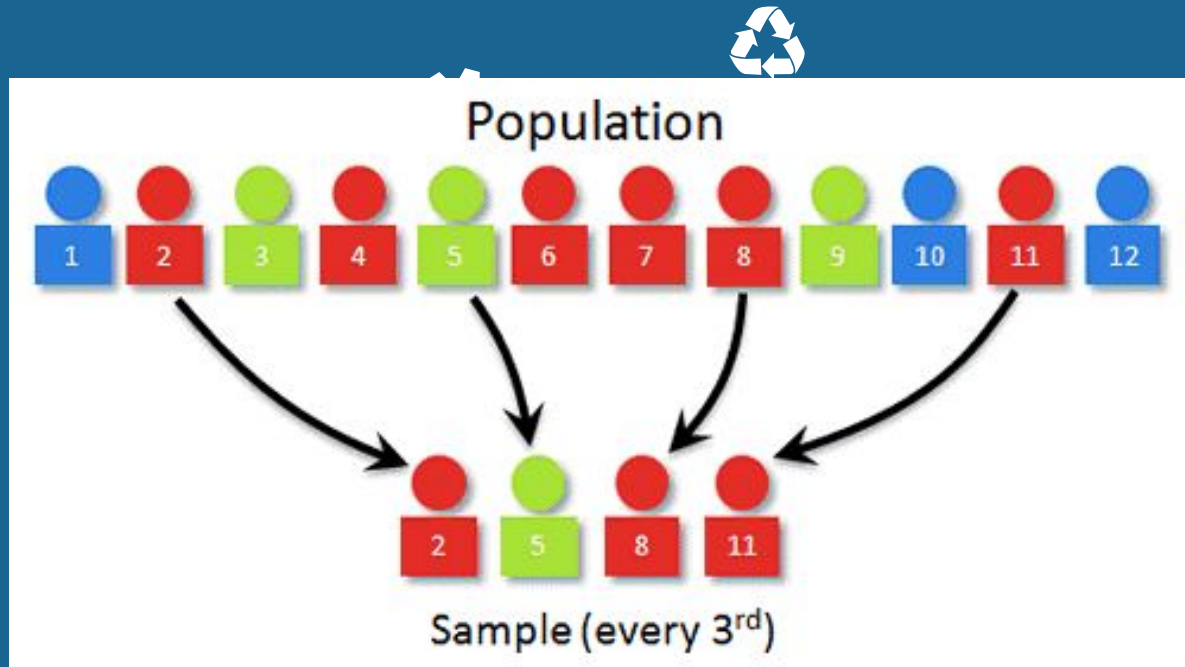


In simple random sampling we randomly choose a member from the population. In this method every member and set of member has an equal chance of being selected in the sample. To conduct this type of sampling, you can use tools like random number generator or other technique that entirely based on chance (SRS/SRSWOR/SRSWR).

Example:

Suppose you want to select a sample of 100 employee of company X from the population of 1000 employee from the company database. So by simple random sampling method you first assign a number to each employee from 1 to 1000 in company database and then randomly draw a 100 employee from the population by using random number generator.

Systematic Random Sampling

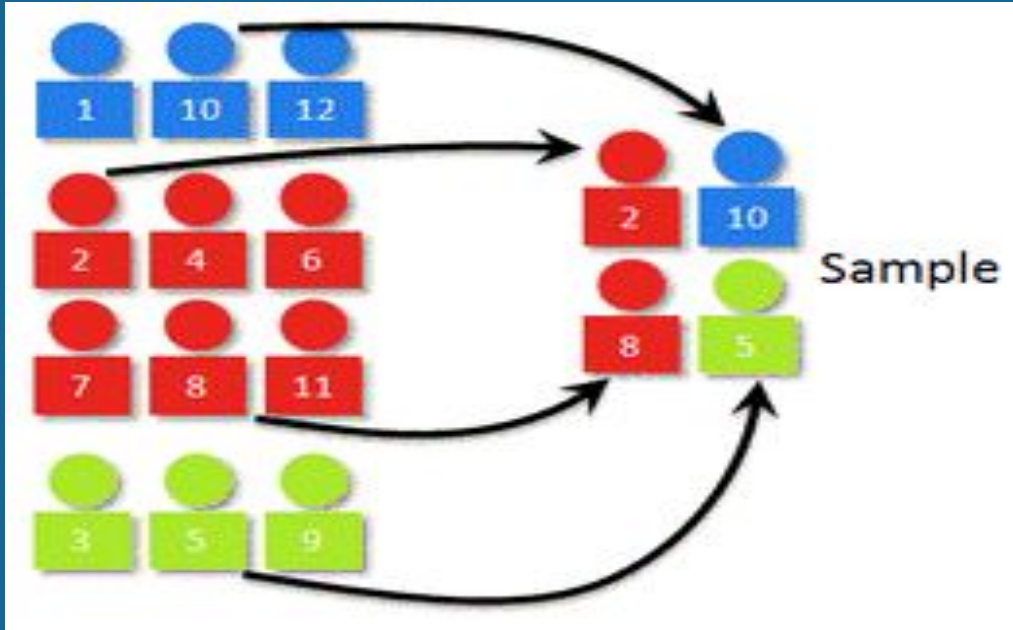


Systematic random sampling is similar to simple random sampling, but it's usually easier to conduct. In this method, we put a member of the population in some order and a starting point is chosen as random. The every "nth" member is selected to be in a sample.

Example:

Let's take the previous example of 1000 employees we want to select 100 samples. From this method, first we arrange the employees' names according to the alphabetical order from A to Z. Now to choose every nth member, we do $1000/100=10$. From the first 10 numbers, you randomly select a number, suppose the number is 5. Now from 5, we select every 10th employee as (5, 15, 25, and so on) and you end with the sample of 100 employees.

Stratified Random Sampling



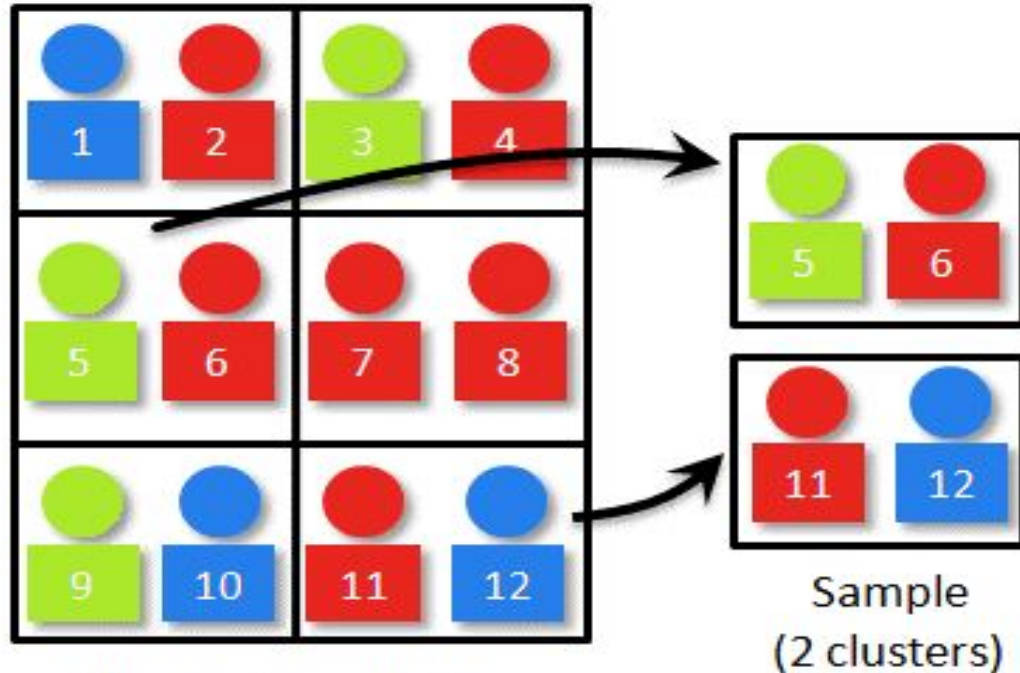
Stratified sampling is appropriate, when population has mixed characteristics and you want to ensure that every characteristics is proportionally represented in sample. In stratified random sampling we first divide the population into groups then from each group we select members randomly.

Example:

The company has 800 female employees and 200 male employees and we want to select 100 sample from the population, you want to ensure that the sample reflect the gender balance of the company. So first we split the population based on there gender, then we use random sampling and select the members randomly from each group, selecting 80 female and 20 male give you a representative sample of 100



Cluster Random Sampling

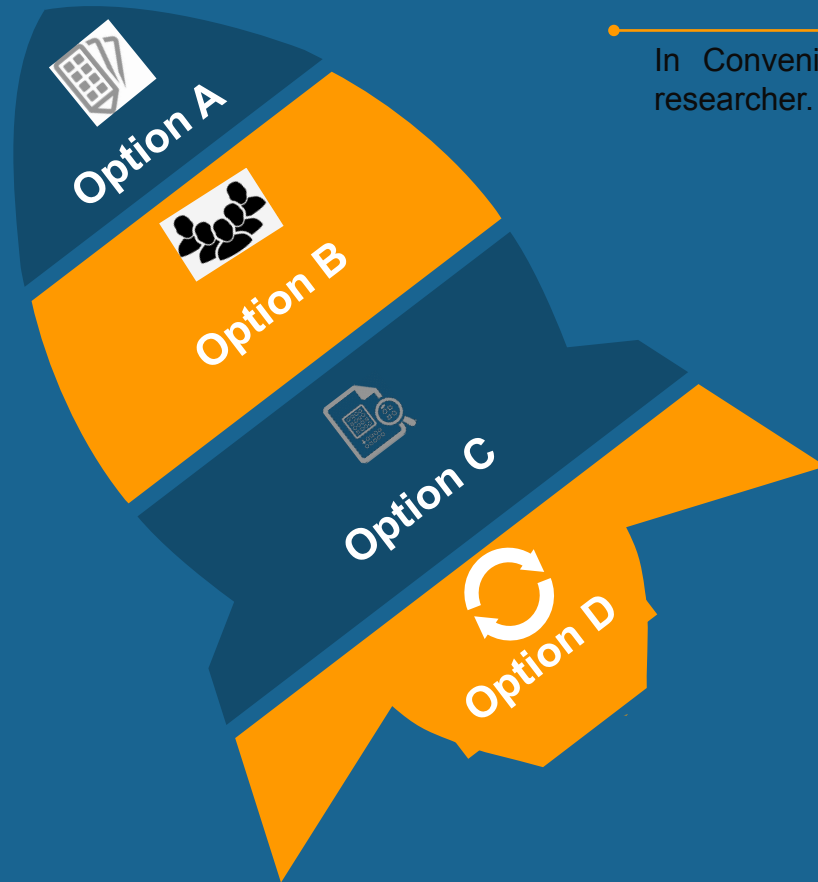


In cluster random sampling we first divide the population into groups and then we randomly select the group from all the groups.

Example:

Consider a scenario where an organization is looking to survey the performance of laptops across India. They can divide the entire country's population into cities (clusters) and select further towns with the highest population and also filter those using laptop devices.

Non-Probability Sampling



Convenience sampling



In Convenience sampling Include the respondents/member who are easy to reach for researcher.

Purposive sampling



In purposive sampling we select a sample based on the purpose of the research. The researcher select the sample by using their expertise and knowledge.

Voluntary Response Sampling



Voluntary response sampling is based on the ease of access. In Voluntary response sampling members volunteer themselves instead of researcher selecting the participants and directly contacting them.

Snowball Sampling



In Snowball sampling we recruit the participants via research participants for test or study. It is used where it's hard to find the potential population for research.

Population Sampling

Population/Sample	Term	Notation	Formula
Population ($X_1, X_2, X_3, \dots, X_N$)	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^N X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$
Sample ($X_1, X_2, X_3, \dots, X_n$) (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	\bar{X}	$\frac{\sum_{i=1}^n X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$



Analysing or testing entire population is impossible and also a cost effective & time taking. To Save our money and time we use the subset of the entire population called sample.

Population sampling is the process of selecting a subset of the objects that is representative of the entire population. The sample must have sufficient size of objects to warrant statistical analysis.

Population sampling must be perform correctly since errors can lead to inaccurate and misleading result.



Central Limit Theorem

Central limit theorem states that when plotting a sampling distribution of mean, the mean of sample mean is equal to the population mean. And then sampling distribution approaches to normal distribution with variance equal to $\frac{\sigma}{n}$.

In simple terms CLT say that no matter how your population data is distributed, the sample data always follows a normal distribution curve.

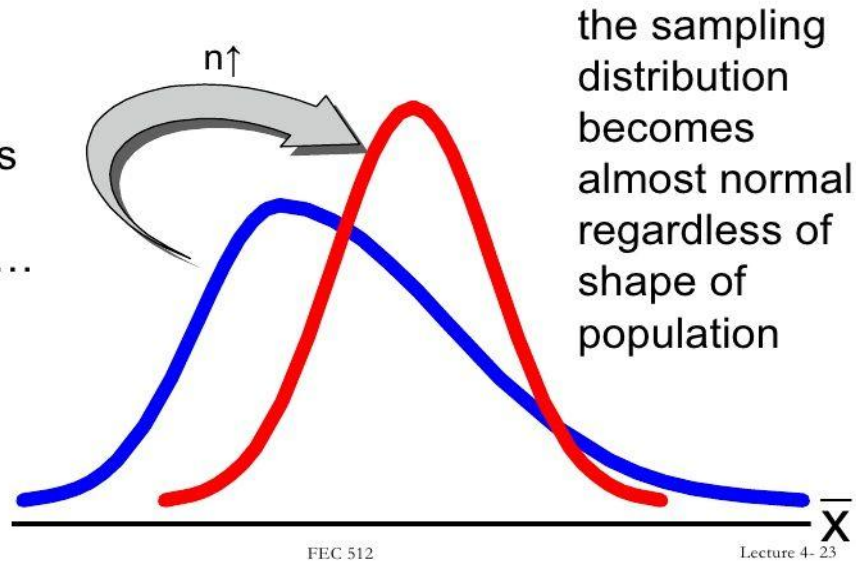
Some important points:

- The sample size must be greater than 30 “ $n > 30$ ” to follow a normal distribution.
- Sampling distribution mean = population mean.

Example

Central Limit Theorem

As the sample size gets large enough...

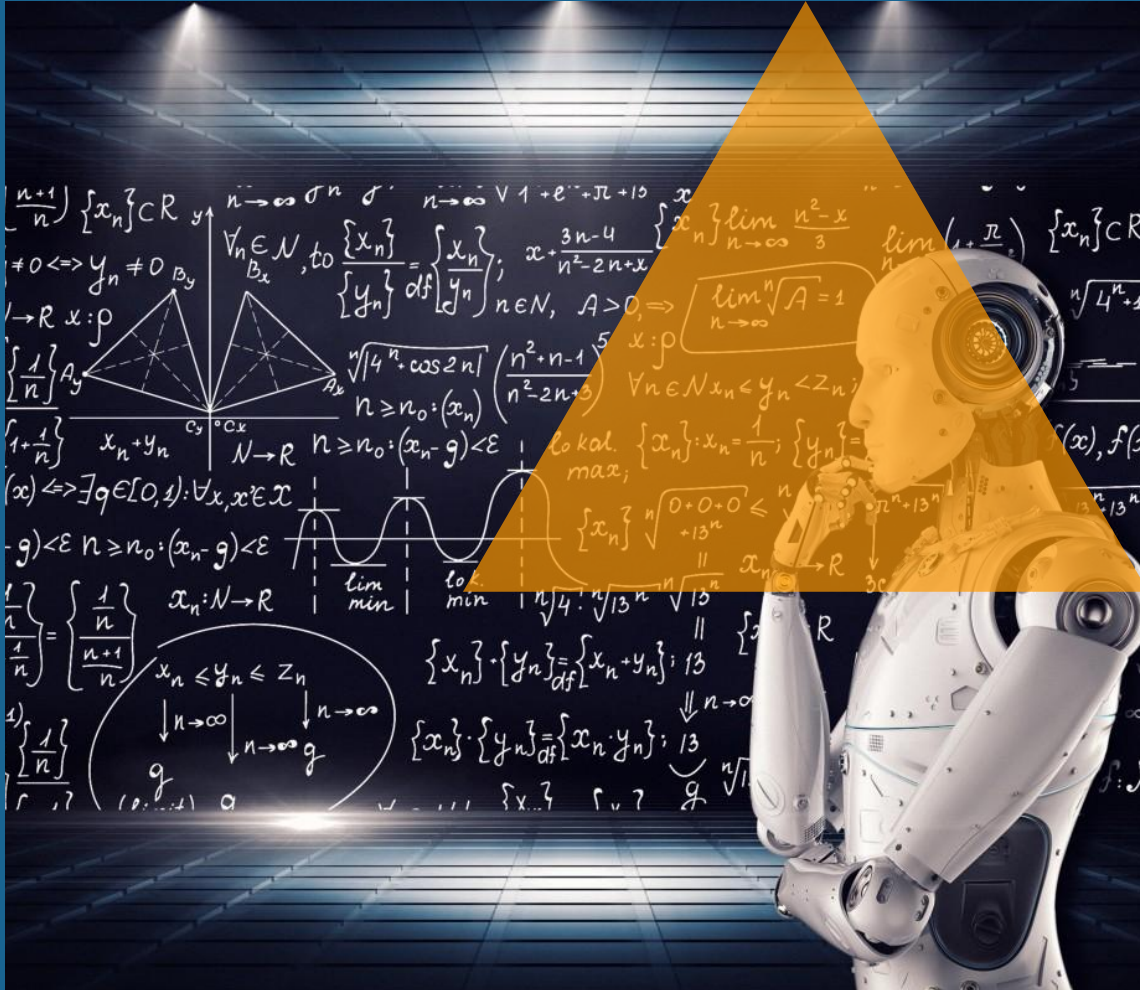


Suppose you want to measure of height of statistics department student of each section there are 15 section in the department and in each section there is approx. 100 students. Our task is to measure average height of students in statistics department.

Now by central limit theorem how we calculate the average height of students in statistics department.

- First, draw groups of students at random from the each class. We will call this a sample. We'll draw multiple samples, each consisting of 30 students.
- Calculate the individual mean of these sample.
- Calculate the mean of these sample mean.
- This value will give us the approximate mean height of the students in the statistics department.
- Additionally, the histogram of the sample mean heights of students will resemble a bell curve (or normal distribution).

Interval



Confidence interval is a type of interval estimate from sampling distribution which gives a range of values in which the population statistic may lie.

$$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

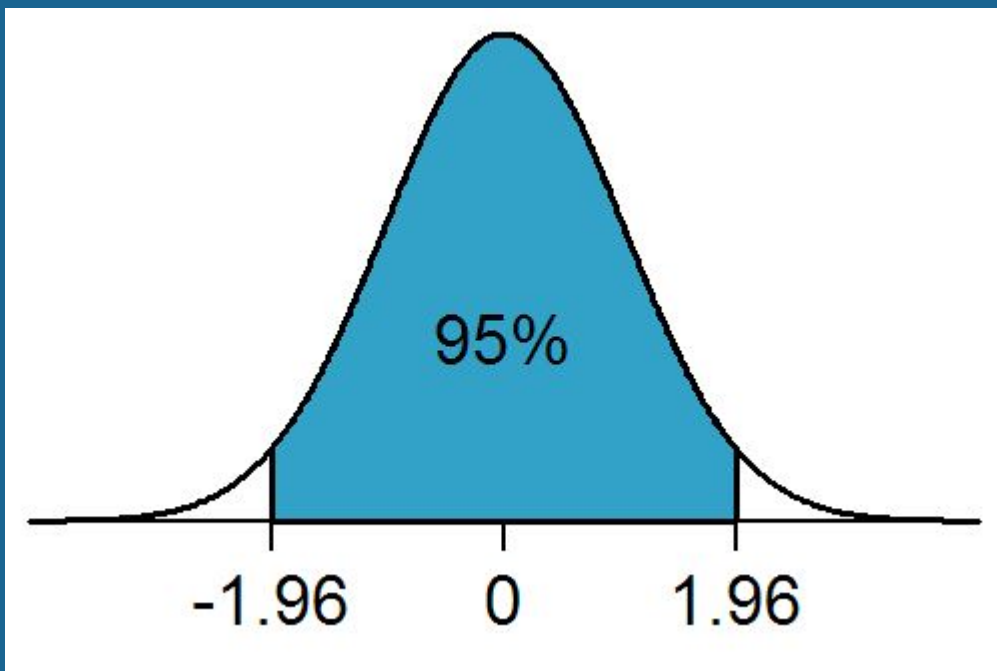
where \bar{X} = the sample mean

σ = the population standard deviation

$Z_{\frac{\alpha}{2}}$ = the Z value for the desired confidence level α (obtained from an Area Under the Normal Curve table)

- ❖ We mainly calculate confidence interval on 95% or on 99%.
- ❖ We interpret the confidence interval result as 95% of the interval estimate will contain the population statistic.
- ❖ Take different confidence interval for different sample mean.

Example



Confidence interval for sample mean 0.

Calculate the 95% confidence interval for a sample mean of 40 and sample standard deviation of 40 with sample size equal to 100.

We know, z-value for 95% C.I is 1.96. Hence, Confidence Interval (C.I) is calculated as:

$$C.I = [\bar{x} - (z \cdot s / \sqrt{n}), \bar{x} + (z \cdot s / \sqrt{n})]$$

$$C.I = [40 - (1.96 \cdot 40 / 10), 40 + (1.96 \cdot 40 / 10)]$$

$$C.I = [32.16, 47.84]$$