Business Statistics

Agenda

- Introduction to Statistics
- Types of Data & Statistical Analysis
- Population and Samples
- Type of Sampling Technique
- Measures of Central Tendency and Dispersion
- Probability
- Probability Distribution
- Covariance & Correlation
- Hypothesis Testing

Introduction to Statistics

Statistics

Basically, the statistical analysis is meant to collect and study the information available in large quantities.

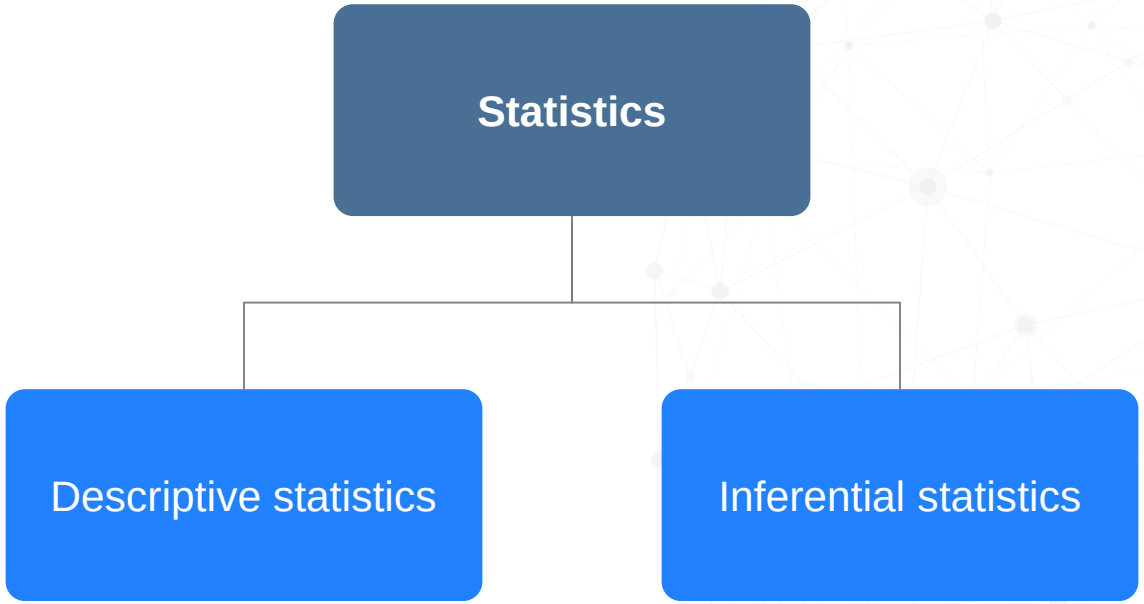
Statistics is a branch of mathematics, where computation is done over a bulk of data using charts, tables, graphs, etc.

The data collected for analysis here is called measurements. Now, if we have to measure the data based on a scenario, a sample is taken out of a population.

Then the analysis or calculation is done for the following measurement.

Types of Data & Statistical Analysis

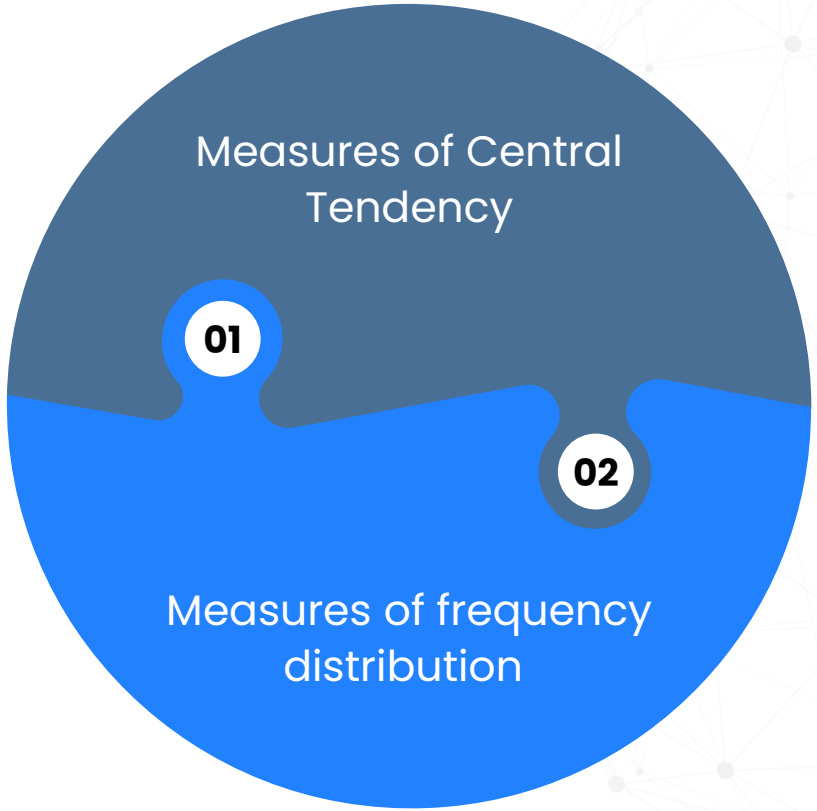
Type of Statistics



Descriptive statistics

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

Descriptive statistics



Inferential Statistics

A set of method that is used to draw a conclusion about the characteristics of population based on the sample of the data

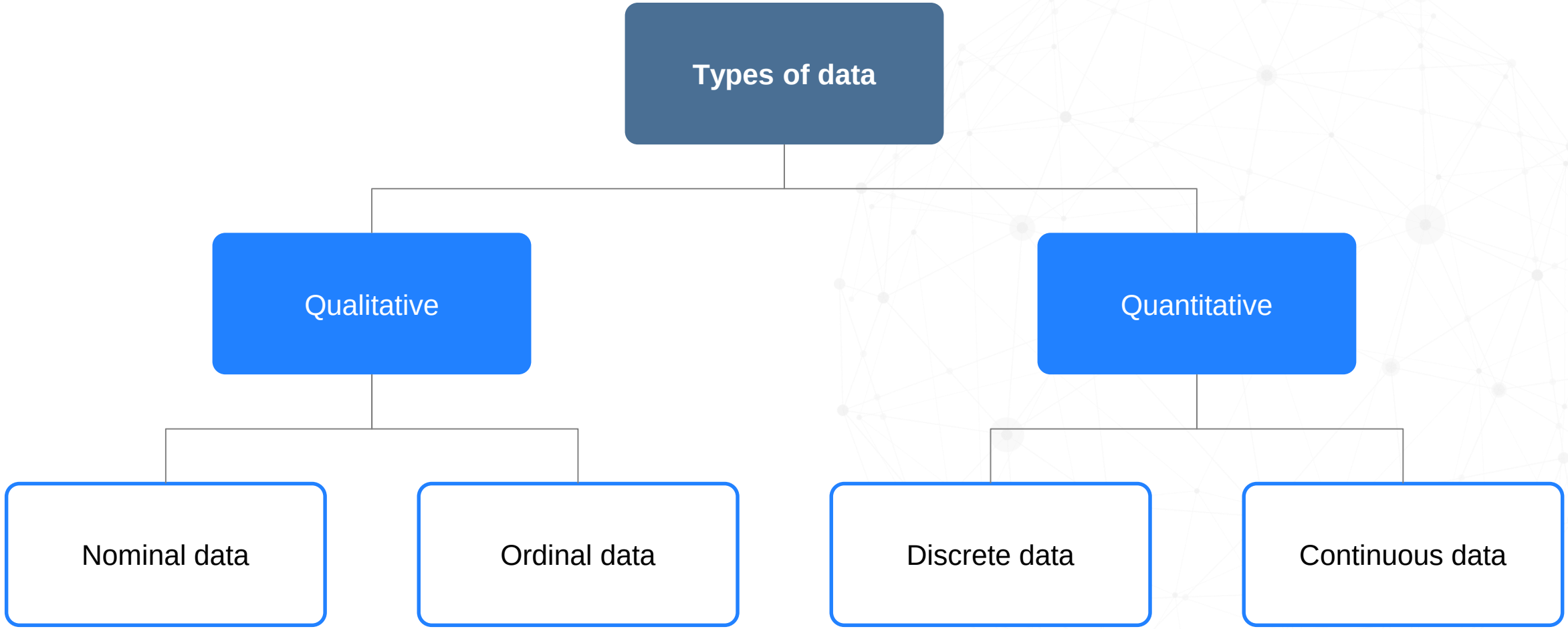
Used to find the population parameter when you have no initial number to start with

The two main areas of inferential statistics

Estimating parameter

Hypothesis Testing

Types of Data

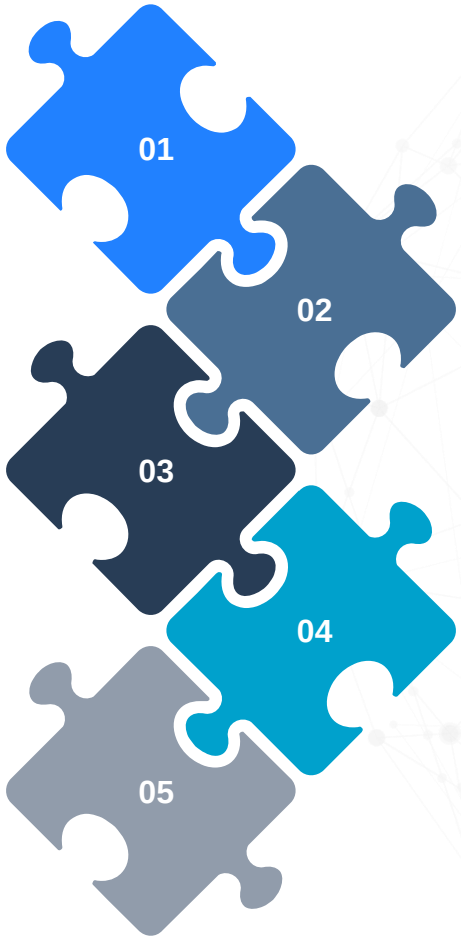


Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories.

The categorical information involves categorical variables that describe the features such as a person's gender, hometown etc.

Examples of the categorical data are birthdate, favorite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.



Qualitative data are not numerical.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense.

Nominal data

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value.

Nominal data is also called the nominal scale. It cannot be ordered and measured.

But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated.

These data are visually represented using the pie charts.

Ordinal data/variable

The significant feature of the nominal data is that the difference between the data values is not determined.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualization tools.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense.

This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The information may be expressed using tables in which each row in the table shows the distinct category.

Quantitative data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many).

Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on.

The quantitative data can be classified into two different types based on the data Sets.

The two different classifications of numerical data are discrete data and continuous data.

Discrete Data and Continuous Data

Discrete Data

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

Example:

Number of students in the class

Continuous Data

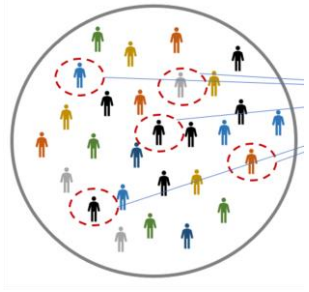
Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

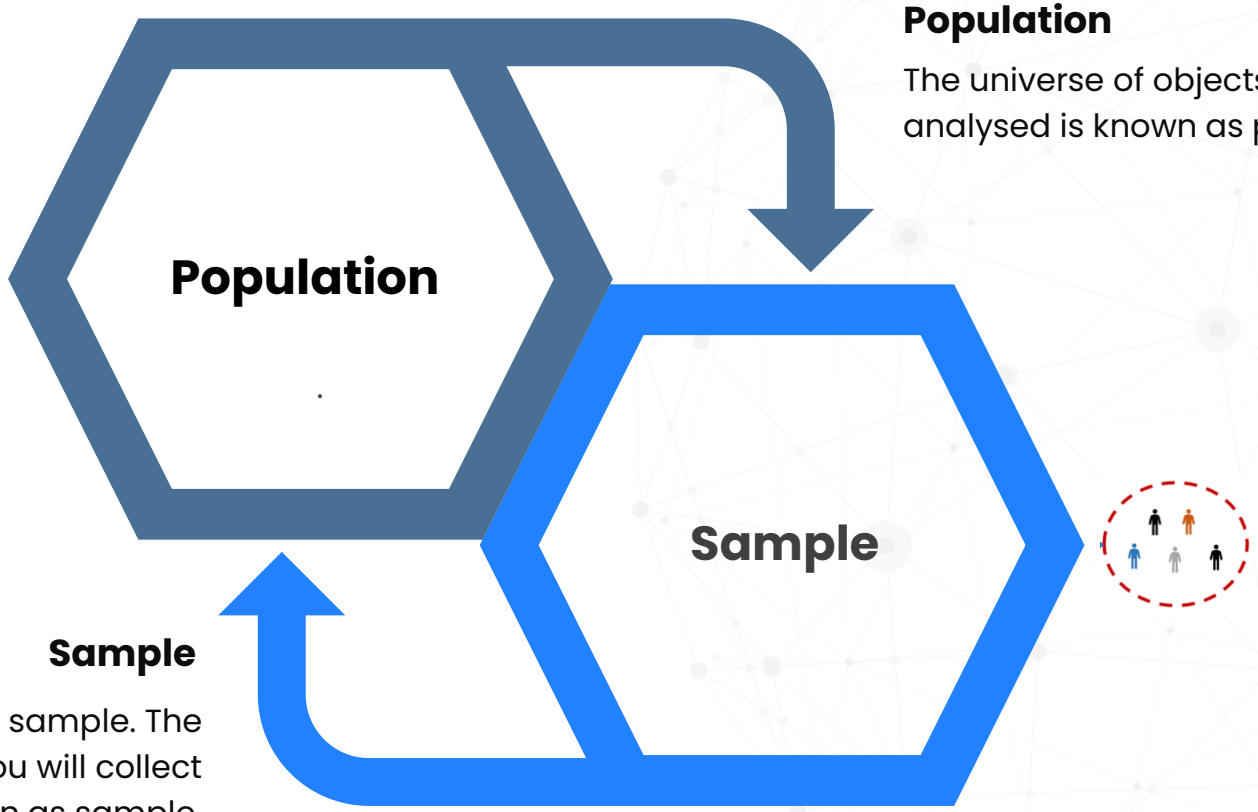
Example:

Temperature range

Population and Sample

Population And Sample





Population

The universe of objects that is required to be analysed is known as population.

Sample

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

Probability Sampling and Non-Probability Sampling

Probability sampling

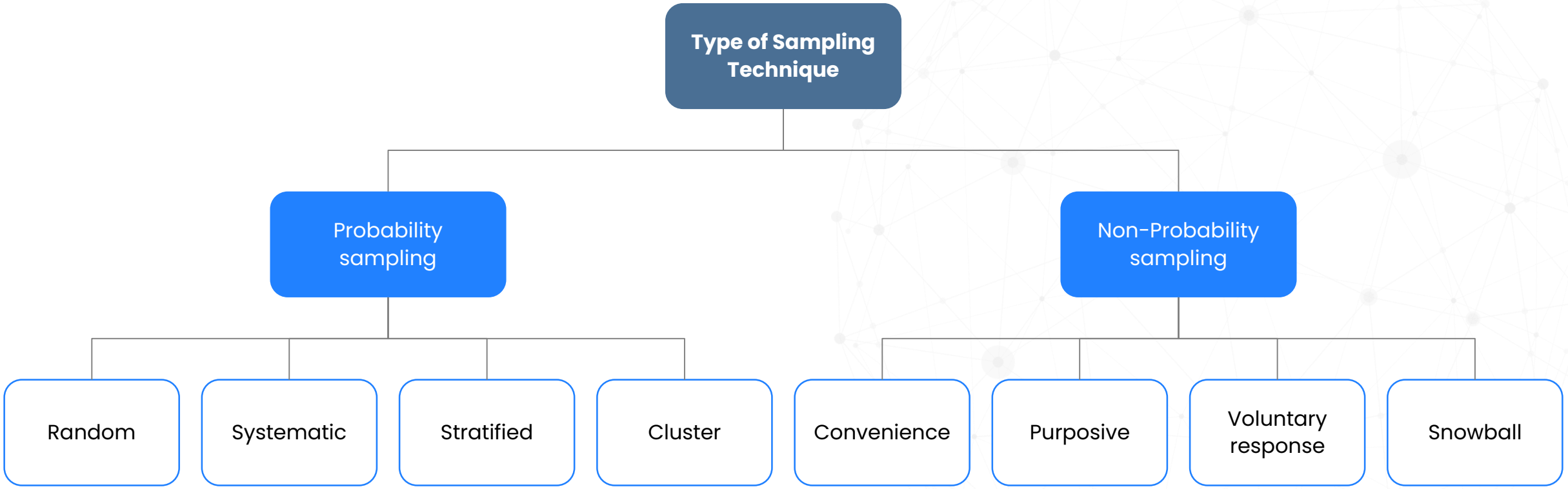
- Every member of the population has an equal chance of being selected
- Examples:
  - Simple random sampling
  - Stratified random sampling
  - Cluster sampling
  - Systematic sampling

Non-probability sampling

- Samples are selected on the basis of judgment or the convenience of accessing data
- Largely depends on a researcher's sample selection skills
- Examples:
  - Convenience sampling
  - Purposive sampling
  - Voluntary response sampling
  - Snowball sampling

Type of Sampling Technique

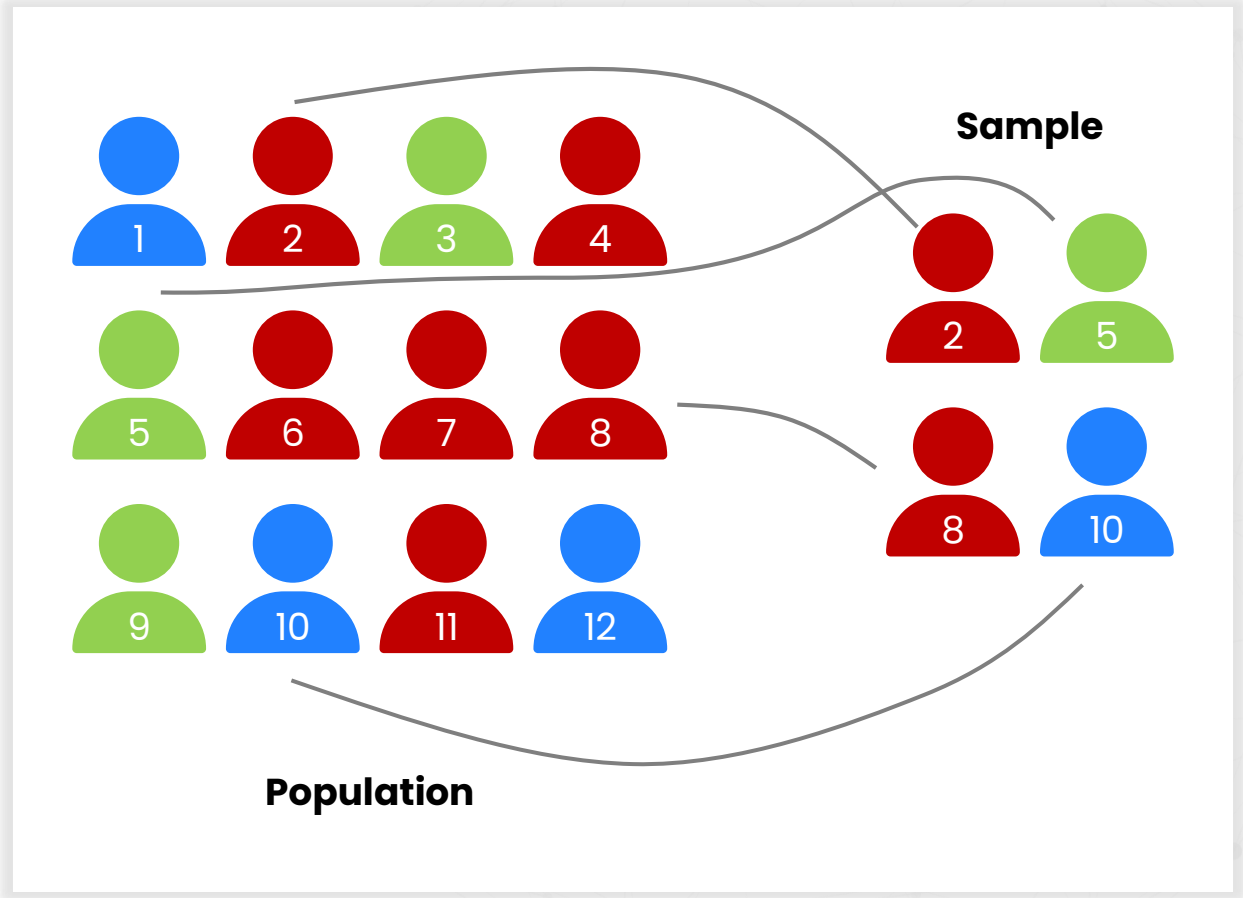
Type of Sampling Technique



Random Sampling

Randomly choose a member from the population

Every member and set of member has an equal chance of being selected.



Systematic Sampling

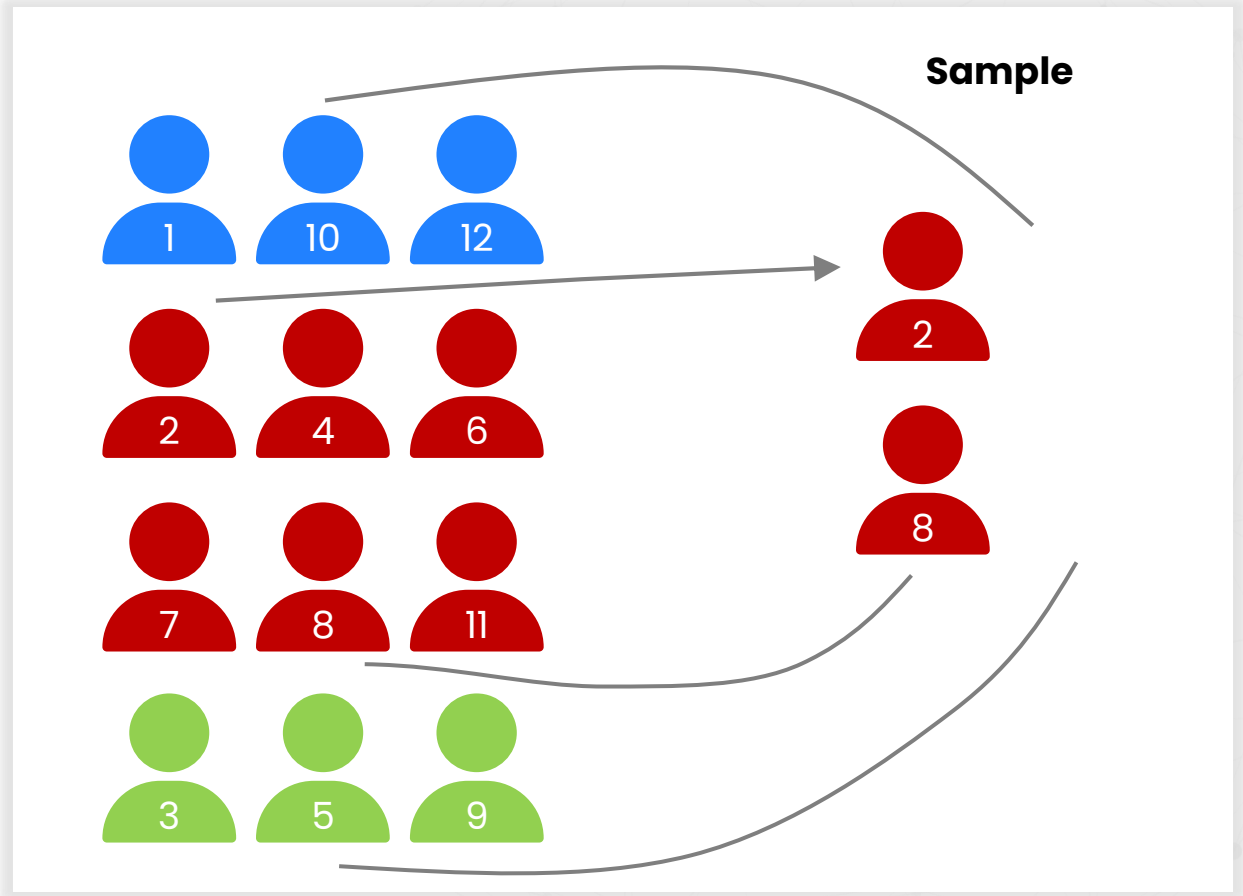
Similar to simple random sampling

Put a member of the population in some order and a starting point is choose as random the every “nth” member is selected to be in a sample.

Stratified Sampling

First divide the population into groups

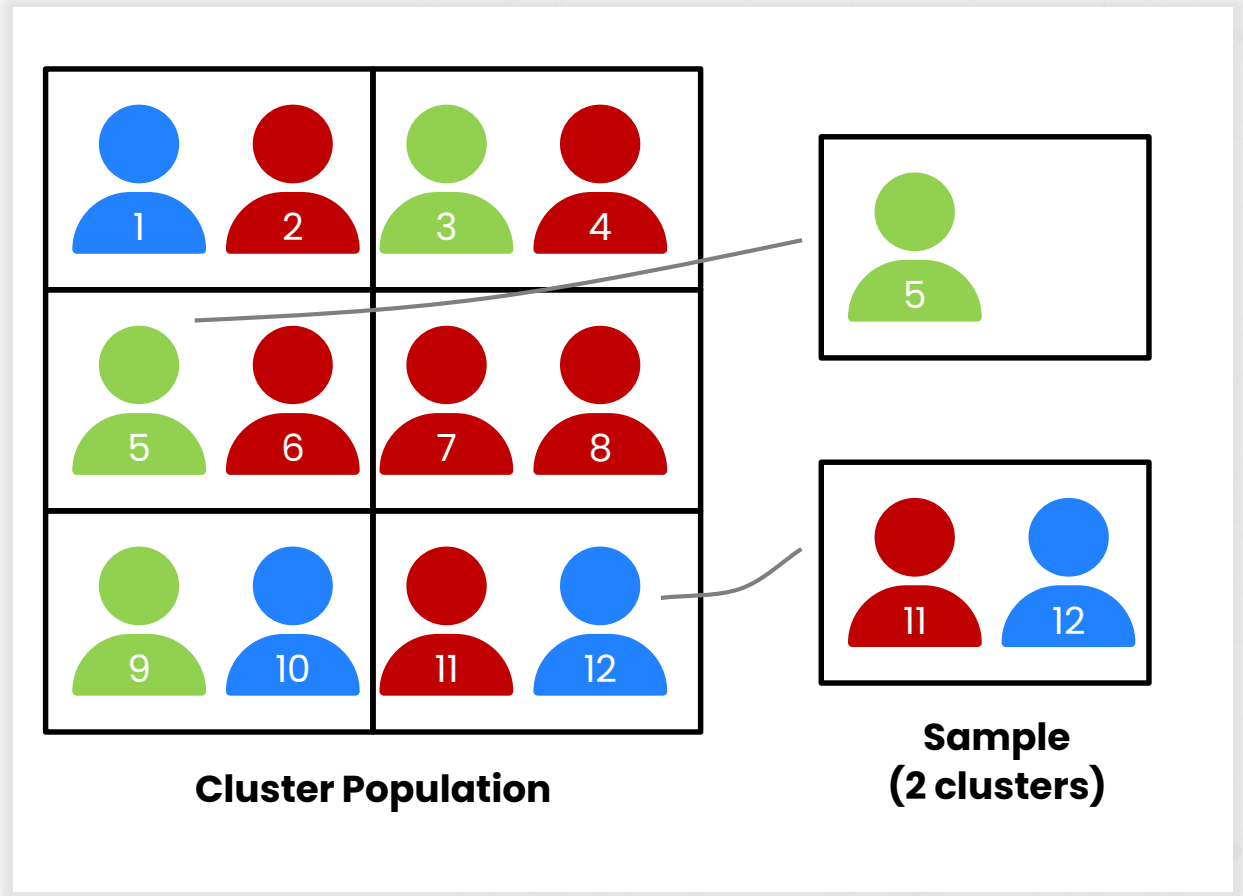
Then from each group we select members randomly.



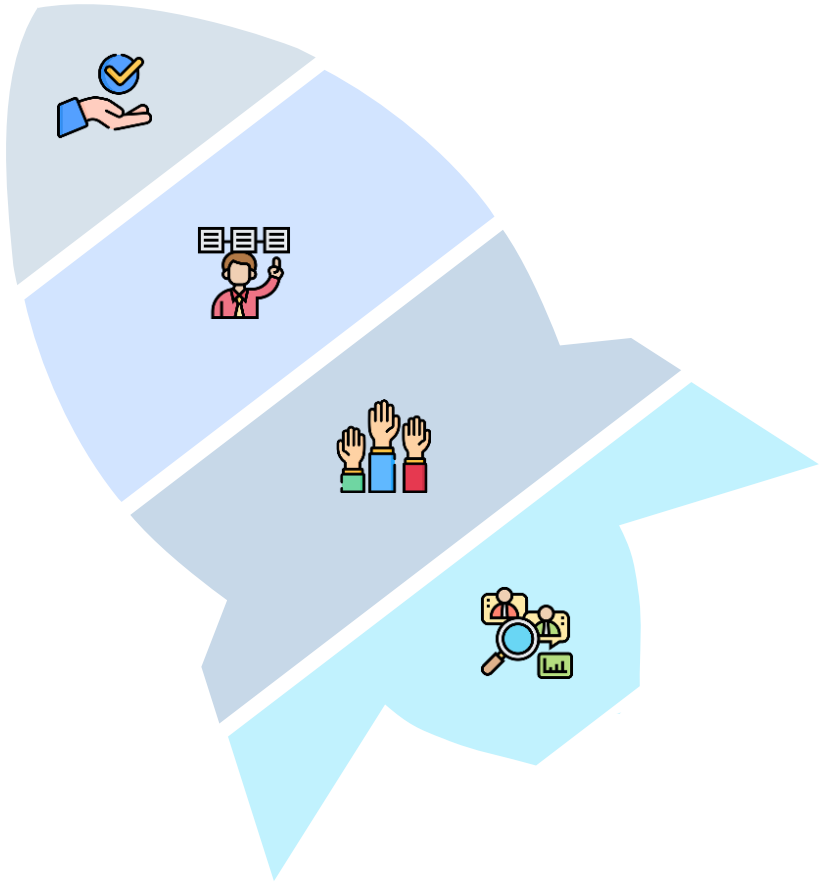
Cluster Random Sampling

Divide the population into groups

Then randomly select the group from all the groups



Non-Probability Sampling



Convenience sampling

- Include the respondents/member who are easy to reach for researcher.

Purposive sampling

- Select a sample based on the purpose of the research.
- Researcher select the sample by using their expertise and knowledge.

Voluntary Response Sampling

- Based on the ease of access.
- Members volunteer themselves instead of researcher selecting the participants and directly contacting them.

Snowball Sampling

- Recruit the participants via research participants for test or study
- Used where it's hard to find the potential population for research

Population Sampling

Analysing or testing entire population is impossible and also a cost & 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.

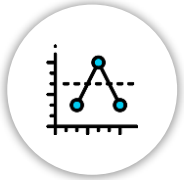
Must be perform correctly since errors can lead to inaccurate and misleading result.

Population/sample	Term	Notation	Formula
Population ($X_1, X_2, X_3, ..., X_N$)	Population size	N	Number of items /elements in the population
	Population mean	μ	
	Population variance	σ^2	
Sample ($X_1, X_2, X_3, ..., X_N$) (Sample of population)	Sample size	N	Number of items/elements in the sample
	Sample Mean		
	Sample variance	S^2	

Measures of Central Tendency and Dispersion

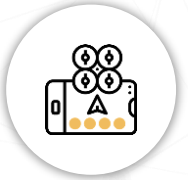
Measures of Central Tendency

- Measures the center value of the dataset.
- Give us idea about the concentration of the value in the central part of the distribution.



Mean / Average

Median



Mode

Mean/Average

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.

Mean/Average

Add/Sum each number/observation present in the dataset.

Calculate the total number present in the dataset.

Divide sum of observation to the total number of observation.

Formula

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

Median

01 The middle number

02 Found by ordering all data points and picking out the one in the middle

03 If there are two middle numbers, taking the mean of those two numbers

Median

Arrange the observation is ascending order

Number of observations (n) is odd

The median is the middle value, which is at position

$$\left(\frac{n+1}{2}\right)$$

Number of observations (n) is even

The median is the average of the two middle values

1. Find the value at position $\left(\frac{n}{2}\right)$
2. Find the value at position $\left(\frac{n}{2}\right)+1$
3. Find the average of the two values to get the median

Mode

- The value which occur most frequently in the set of the observation
- Can have more than one mode as Uni-modal, Bi-modal, Multi modal

Steps for finding the Mode

- It's very easy to find mode of any observation
- Take the Most frequent value present in the dataset.

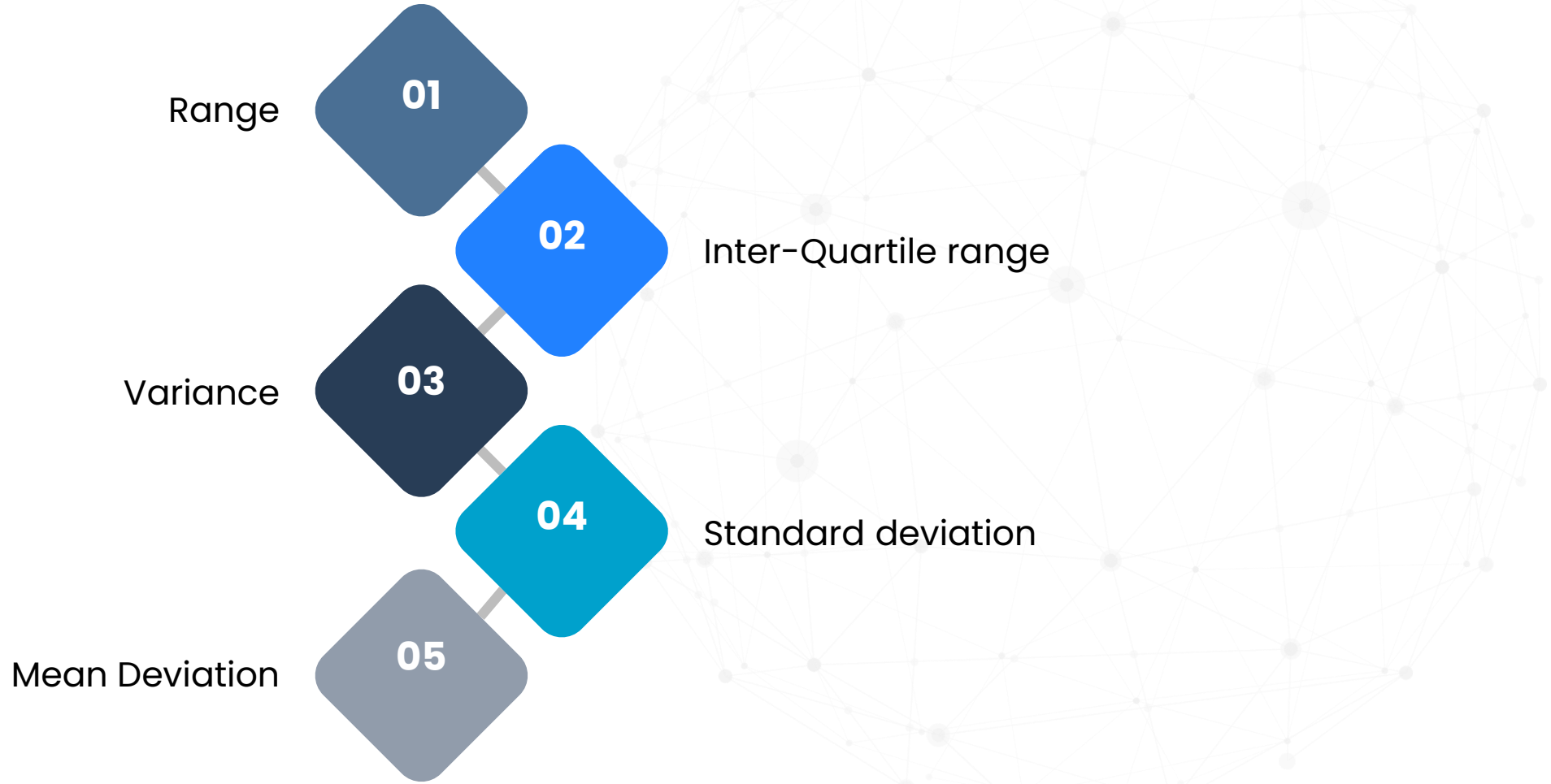
Special Cases

- If the maximum number of frequency repeated
- If the maximum frequency is occurred at the beginning and end of the observation.
- If there is irregularities in the distribution.

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

Measures of Dispersion

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



Range

Simplest measure of dispersion

Measures the difference between highest value and lowest value present in the dataset

Used to construct control chart in quality assurance

Useful when you want to focus on extremes values of the dataset.

The formula of Range is:
Range = Highest value – lowest value

Inter-Quartile Range

Measure the middle 50% of the data

Indicates how the data is dispersed around the mean.

Difference between the third quartile and first quartile value of the dataset.

Helpful to detect the outlier present in the dataset

The formula of IQR is:
IQR=Q3-Q1

Variance

- Measure the dispersion of the data around the mean of the data
- 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.
- Denoted by σ^2

Formula for population variance:

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

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

Formula for sample variance:

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

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

Standard deviation

- Most important and frequently method in measure of dispersion
- Simply the Square root of the variance
- Indicate how far away the datapoints is dispersed from the mean
- 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 data points and $\bar{x}$ is sample mean

Why n-1 & why not n?

- This is actually called as Bessel's correction
- This method corrects the bias in the estimation of the population variance.
- It also partially corrects the bias in the estimation of the population standard deviation
- The idea behind this is that this is a more unbiased measure of variance than the usual definition.

Imagine having a huge bookshelf.

Let say, the total thickness of the first 6 books turns out to be 158mm. This means that the mean thickness of a book based on first 6 samples is 26.3

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 nth object has no freedom to vary. Therefore, degree of freedom is only (n-1) and not n.

Mean Deviation

- Average sum of the absolute values of the deviation from any arbitrary values, e.g. mean, median, mode etc.
- Suggested to calculate from median because its give least value when measured from the median.
- The deviation of an observation xi from the assumed mean
  - A is defined as (xi - A)
- Therefore, the mean deviation can be defined as

$$\textbf{Mean deviation} = \frac{\Sigma \mid X - \mu \mid}{N}$$

Probability

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	Event	Sample Space	Random Experiment
Each performance in a random experiment	Outcome of trial is an event P(E).	Set of all possible outcome in a random experiment P(s).	To perform more than once

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}

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

P(A) = 3/6

P(A) = ½

So probability of getting a even number in single die is ½

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}

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

P(A) = 1/2

P(A) = ½

So probability of getting a head of tossing a fair coin is ½

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..

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

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.

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

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 a 2 or 5

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

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

P(2 or 5) = P (2) + P (5)

= 1/6 + 1/6

= 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(Girl) = 13/30

P (A) = 9/30

P(A∩B) = 5/30

P(Girl or A) = P (Girl) + P(A) – P(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:

- $P\left(\frac{A}{B}\right) = P(A)$
- $P\left(\frac{B}{A}\right) = P(B)$
- $P(A\ \&B) = P(A)\ * P\ (B)\ (Multiplication\ rule\ for\ independent)$

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

Cumulative Probability

Cumulative probability refers to the likelihood that the value of a random variable is within a given range.
For example,
$P(a \leq X \leq b)$,
Where X is a random variable and a and b are the range limits.

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:

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

Conditional Probability

When we know that a particular event B has occurred instead of A, we concentrate our attention on B only and the conditional probability of A given B will be analogously the ratio of probability of that part of A which is included in B to the probability of B

$P(B|A) = P(A \cap B) / P(A)$

Rule of multiplication

$P(A \cap B) = P(A|B) * P(B)$ or

$P(A \cap B) = P(B|A) * P(A)$

Example:- An urn contains 6 red marbles and 4 black marbles. Two marbles are drawn without replacement from the urn. What is the probability that both of the marbles are black?

Solution:

- Let A = the event that the first marble is black; and let B = the event that the second marble is black. We know the following:
- In the beginning, there are 10 marbles in the urn, 4 of which are black. Therefore, $P(A) = 4/10$.
- After the first selection, there are 9 marbles in the urn, 3 of which are black. Therefore, $P(B|A) = 3/9$.
- Therefore, based on the rule of multiplication:

  $P(A \cap B) = P(A)\, P(B|A)$

  $P(A \cap B) = (4/10) * (3/9) = 12/90 = 2/15 = 0.133$

Conditional Probability - Example

Example:- In a group of 100 people, 40 bought sports drinks, 30 purchased snacks, and 20 purchased sport drinks and snacks. If a buyer chosen at random bought a sport drinks, what is the probability that they also bought snacks?

Bayes Theorem

Bayes Theorem is the extension of Conditional probability.

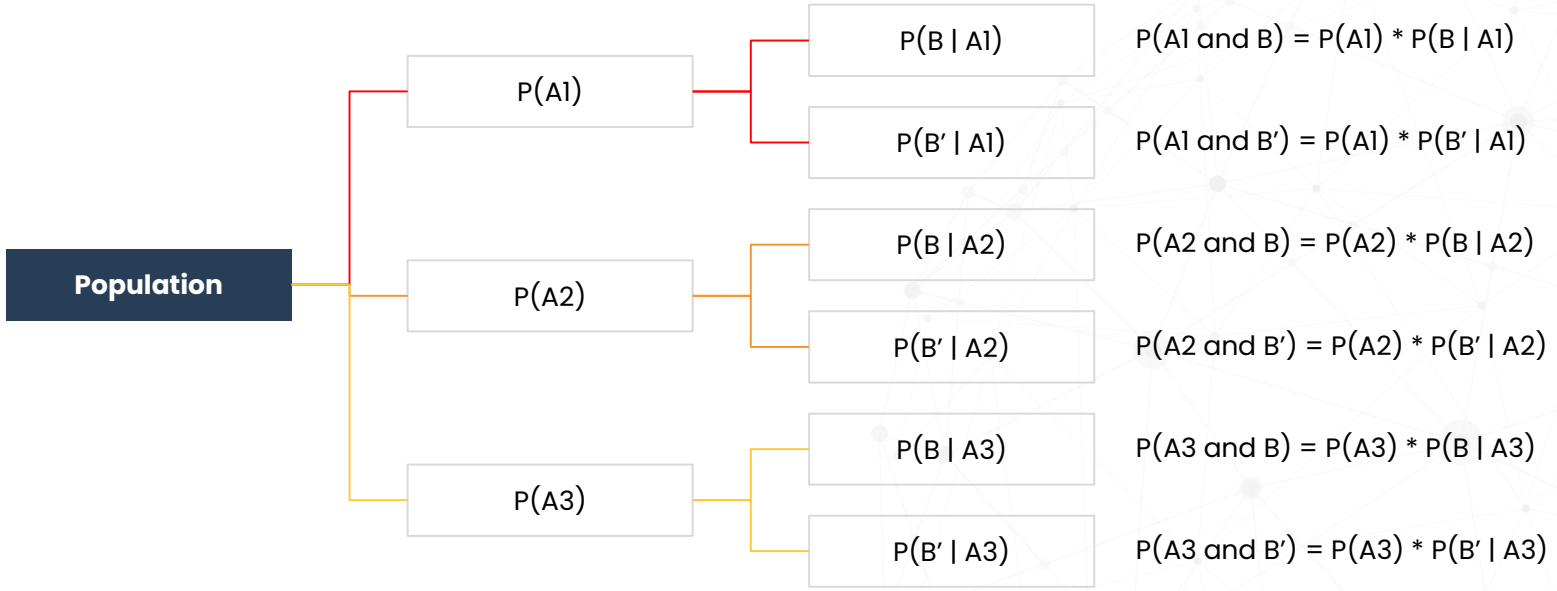
Conditional probability helps us to determine the probability of A given B, denoted by P(A|B).

Bayes' theorem says if we know P(A|B) then we can determine P(B|A), given that P(A) and P(B) are known to us.

Formula: $P(A|B) = P(A) * P(B|A) / P(B)$

Bayes Theorem - Example

Example:- A Bag X contains 4 white & 6 black balls while another Bag Y contains 4 white & 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from Bag X.



Bayes Theorem: To Find reverse probabilities

$P(A1 \mid B) = P(A1) * P(B \mid A1) / P(B)$

P(A1) and P(B) are known as marginal probabilities.
P(B|A1) and P(A1) is given to us.
P(B) can be calculated as
$P(B) = P(A1) * P(B \mid A1) + P(A2) * P(B \mid A2) + P(A3) * P(B \mid A3)$ and also known as total probability

Bayes Theorem: Find reverse probabilities

$P(A1 \mid B') = P(A1) * P(B' \mid A1) / P(B')$

P(B′) can be calculated as

$P(B') = P(A1) * P(B' \mid A1) + P(A2) * P(B' \mid A2) + P(A3) * P(B' \mid A3)$

Probability Distribution

Uniform Distribution

All possible outcomes are equally likely to happen.

- Example: throwing a dice, each number would have the same possibility to happen (1/6)

Two types of uniform distribution

Discrete Uniform

- In layman term, each integer value between a and b has the same possibility to happen
- P(X = x) = 1/n

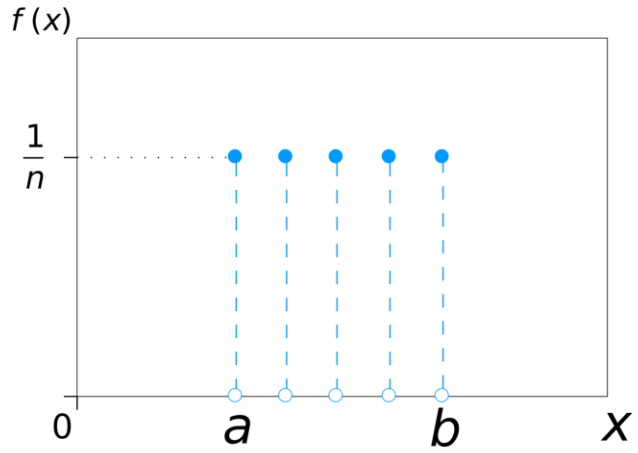
Continuous Uniform

- In layman term, any values (integer or decimal values) between a & b has the same possibility to happen

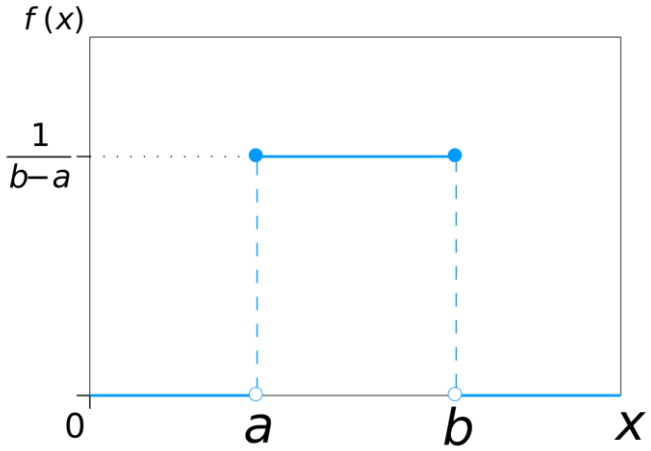
F(x) = {1/(b-a), if a<x<b
0, otherwise}

Where, F(x) = value of density function at any X value
a = lower limit of interval
B = upper limit of interval

Discrete Uniform Distribution



Continuous Uniform Distribution



Uniform Distribution Example

If you are rolling a die once, the probability of each of the possible outcome (i.e. 1, 2, 3, 4, 5, or 6) is even.

$$P\,(X = 1) = \frac{1}{6}$$

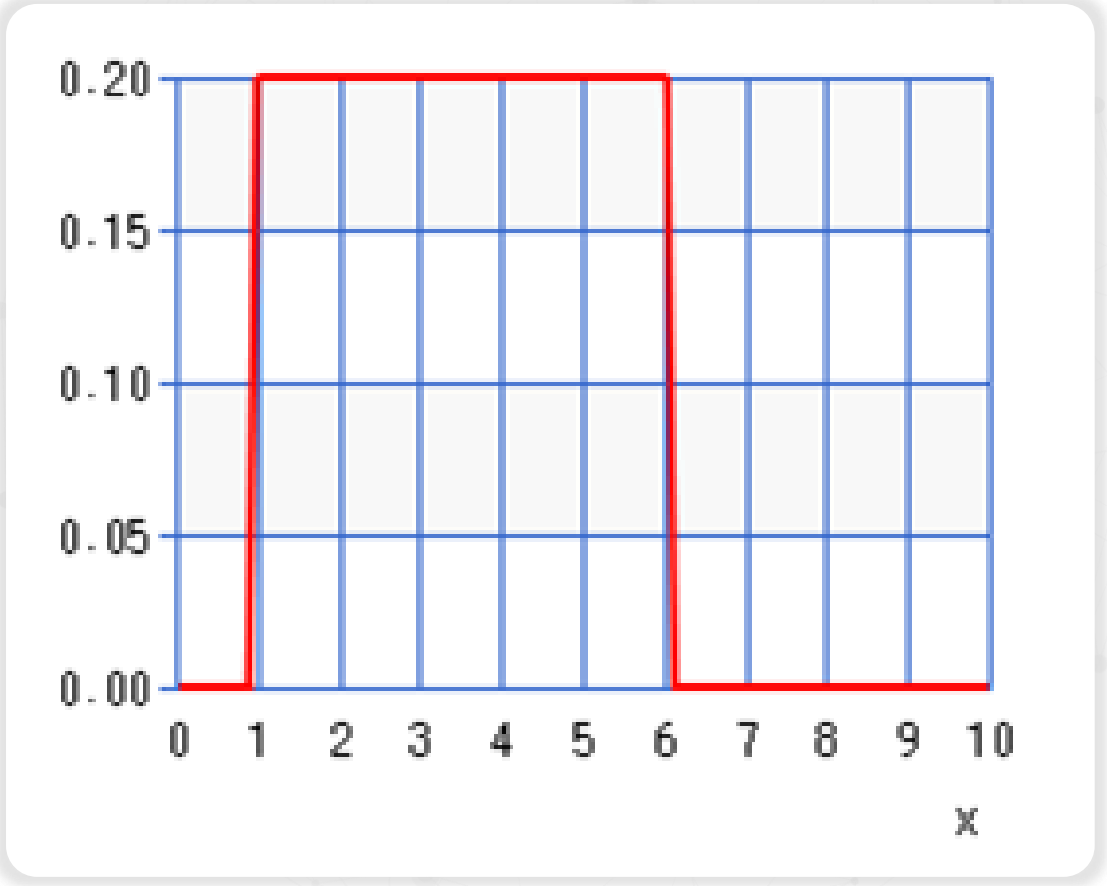
$$P\,(X = 2) = \frac{1}{6}$$

$$P\,(X = 3) = \frac{1}{6}$$

$$P\,(X = 4) = \frac{1}{6}$$

$$P\,(X = 5) = \frac{1}{6}$$

$$P\,(X = 6) = \frac{1}{6}$$



Binomial Distribution

- **N** identical trials are performed where is **n** determined prior to the experiment.
- Trials are independent
- Each trials has two possible outcomes which we call success of failure.
  - Example: tossing a coin
  - The probability of success is same for each trial and is denoted by "p". Thus for each trial
    - **P(Success) = p**
    - **P(failure) = q = 1-p**
  - Random variable (X): the total number of success in "n" trial

Binomial distribution formula

$$P(x) = {}_nC_x p^x (1-p)^{n \quad x}$$

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

Binomial distribution formula

$$P(x) = {}_nC_x p^x (1-p)^{n\ \ x}$$

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

Example

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

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

No. of trial (n) = 10

X = 6

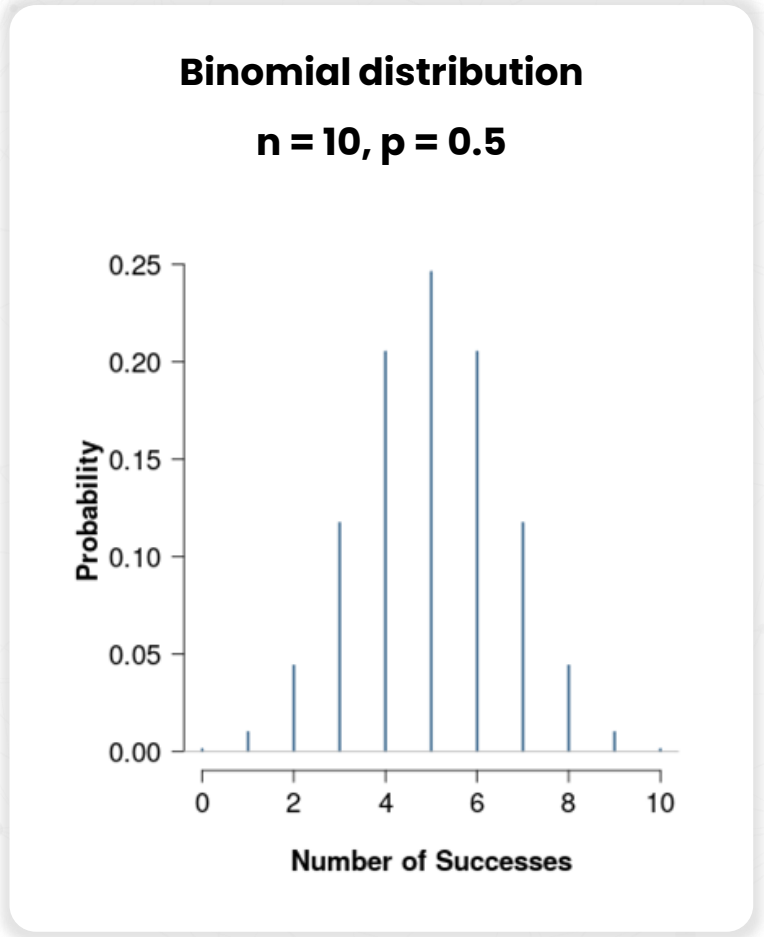
P(success) = probability getting a head = p = 0.5

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

By Binomial distribution:

P(X=6) = $C_6^{10}\ (0.5)^6\ (1\ -0.5)^6$ = 0.205078125

The probability of getting exactly 6 head is 0.205078125



Poisson Distribution

The discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period.

- Let X be the discrete random variable that represents the number of events observed over a given time period.
  - Example: Calls per Hour at a Call Center
- Let λ be the expected value (average) of X
  - e.g. average numbers of calls received in a call center is 10, then λ is 10
- If X follows a Poisson distribution, then the probability of observing x events over the time period is:

$P\ (X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$, for x = 0, 1, 2,... where e is Euler's number (2.71828)

Example – Call Center

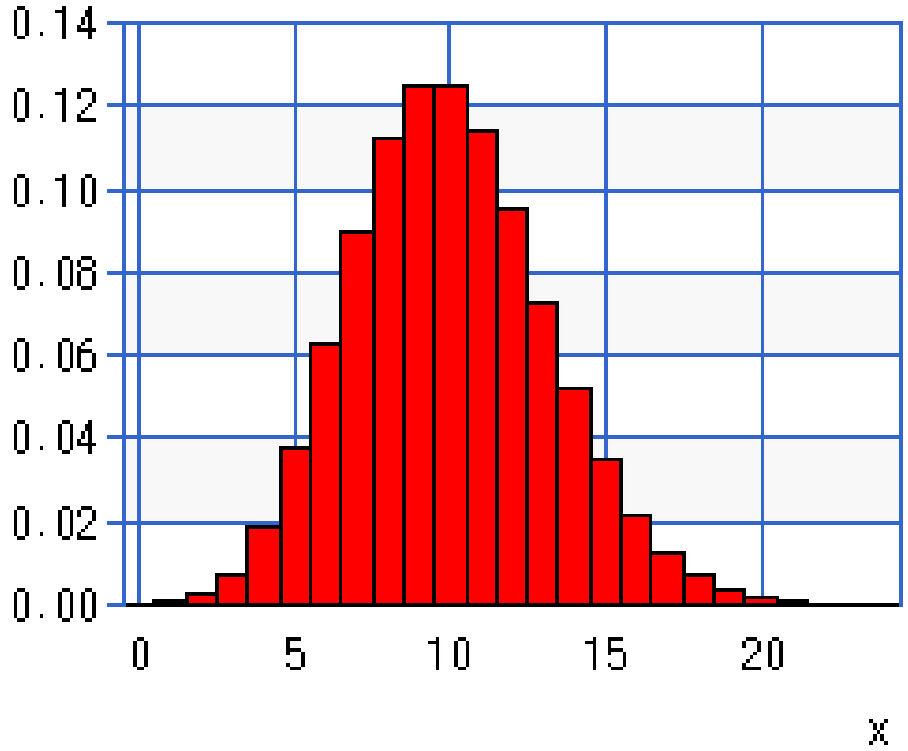
- Average number of calls for a call center is per hour $\lambda = 10$
- Probability of having different number of calls in an hour could be calculated as:

$$P\ (X = 0\ calls) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^0 e^{-10}}{0!} = \frac{1 \cdot 0.000045}{1} = 0.0045\%$$

$$P\ (X = 3\ calls) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^3 e^{-10}}{3!} = 0.75\%$$

$$P\ (X = 5\ calls) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^5 e^{-10}}{5!} = 3.78\%$$

$$P\ (X = 10\ calls) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{10^{10} e^{-10}}{10!} = 12.5\%$$

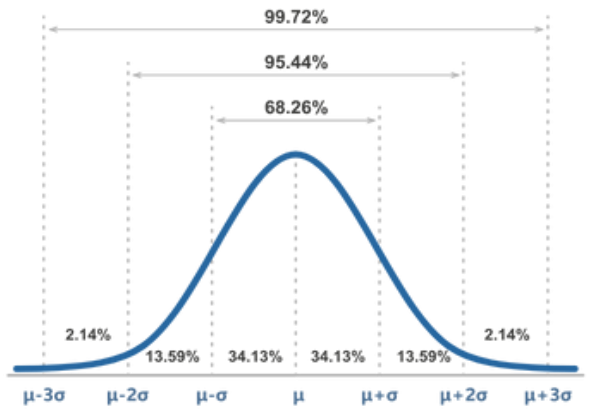


Normal Distribution

- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- Graphically, normal distribution appears as a “bell curve”
- It is important in statistics and is key to the Central Limit Theorem.

The Empirical Rule

For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% of the observations will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations.



$$P(x) = {}_nC_x p^x (1-p)^{n\ \ x}$$

No. of trial (n) = 10

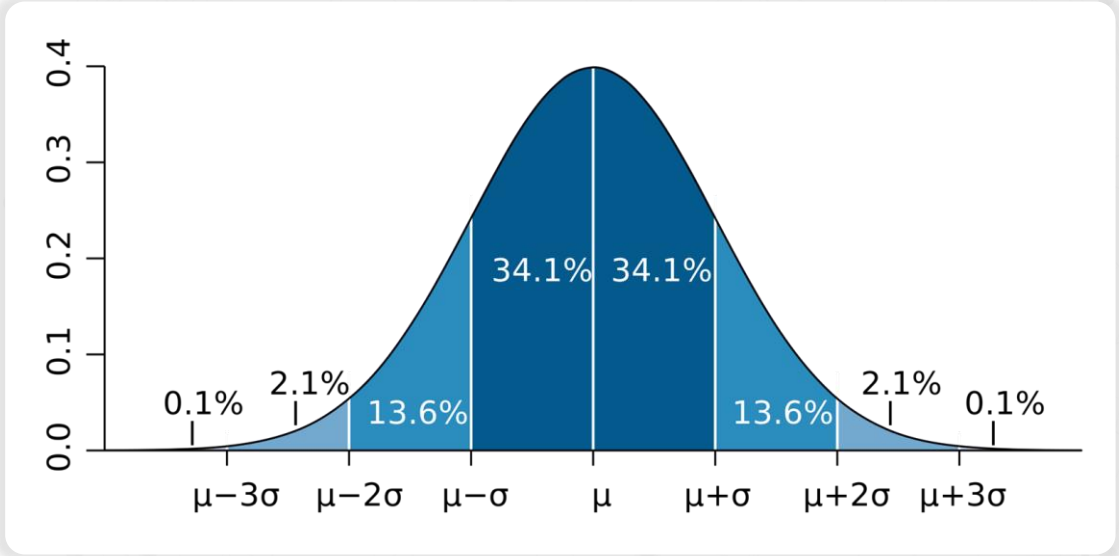
X = 6

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

Normal Distribution (Gaussian distribution)

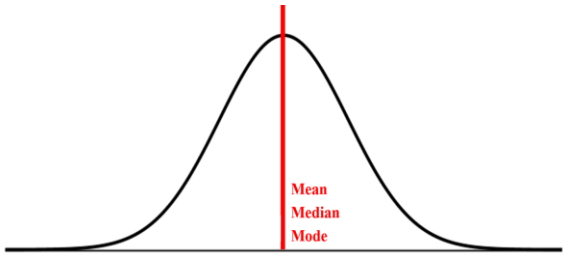
- Bell shape probability distribution
  - Most of the numbers are mid range, will less extreme values
  - Example 1: most people get mediocre score while small number of people get low score or high score
  - Example 2: most people have an average height, but small number of people are extremely tall or short

- Normally distributed data follows the 1-2-3 rule. This rule states that there is a:
  - 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.

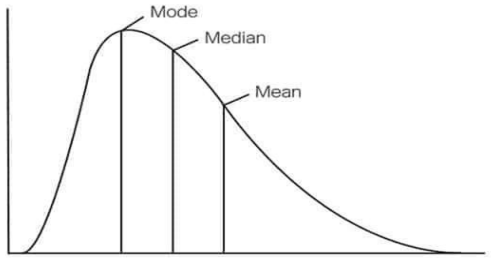


Skewness

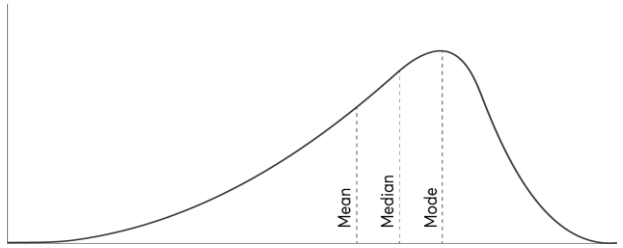
- Skewness measure the distribution of the data. It indicate whether the data is distributed symmetric or not.
- If it is symmetric, then data is normally distributed
- If it is not symmetric, then the data is not normally distributed
- Types of Skewness:
  - Symmetric
  - Positive Skewness
  - Negative Skewness



Normally distributed
(Symmetric)



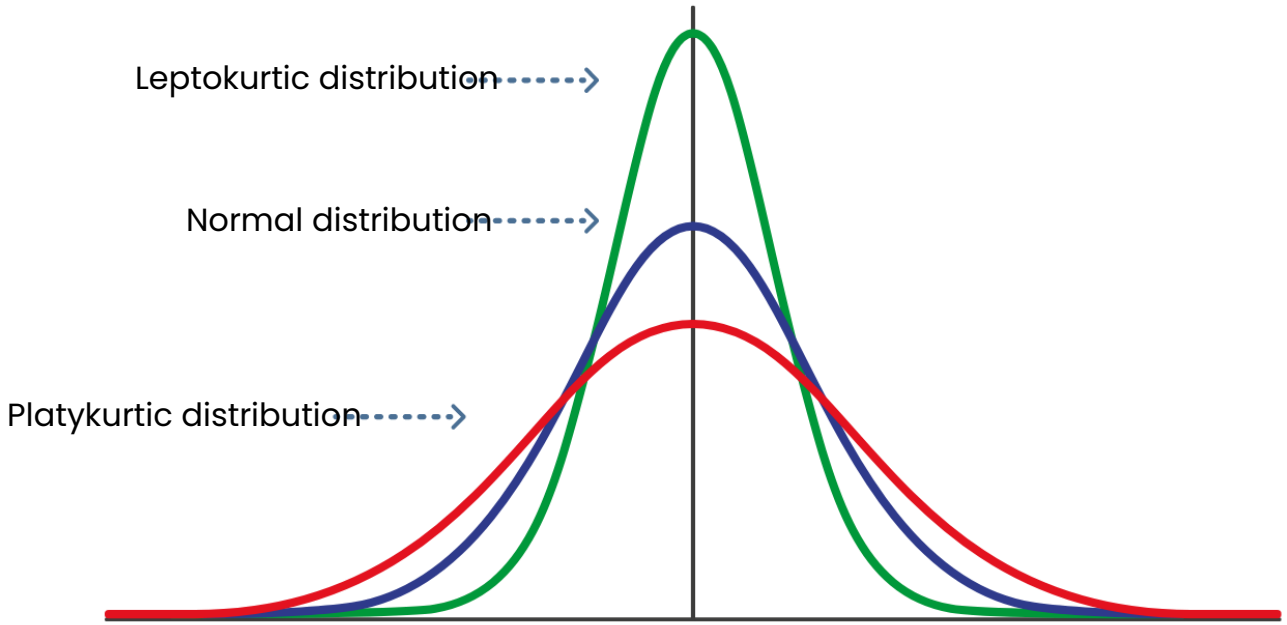
Right-skewed



Left-skewed

Kurtosis

- Kurtosis measure the thickness of the distribution of data.
- The degree of tailed ness of the data is measured by kurtosis.
- It tells us the extent to which the distribution is more or less prone than the normal distribution.
- Types of Kurtosis:
  - Platykurtic
  - Mesokurtic
  - Leptokurtic

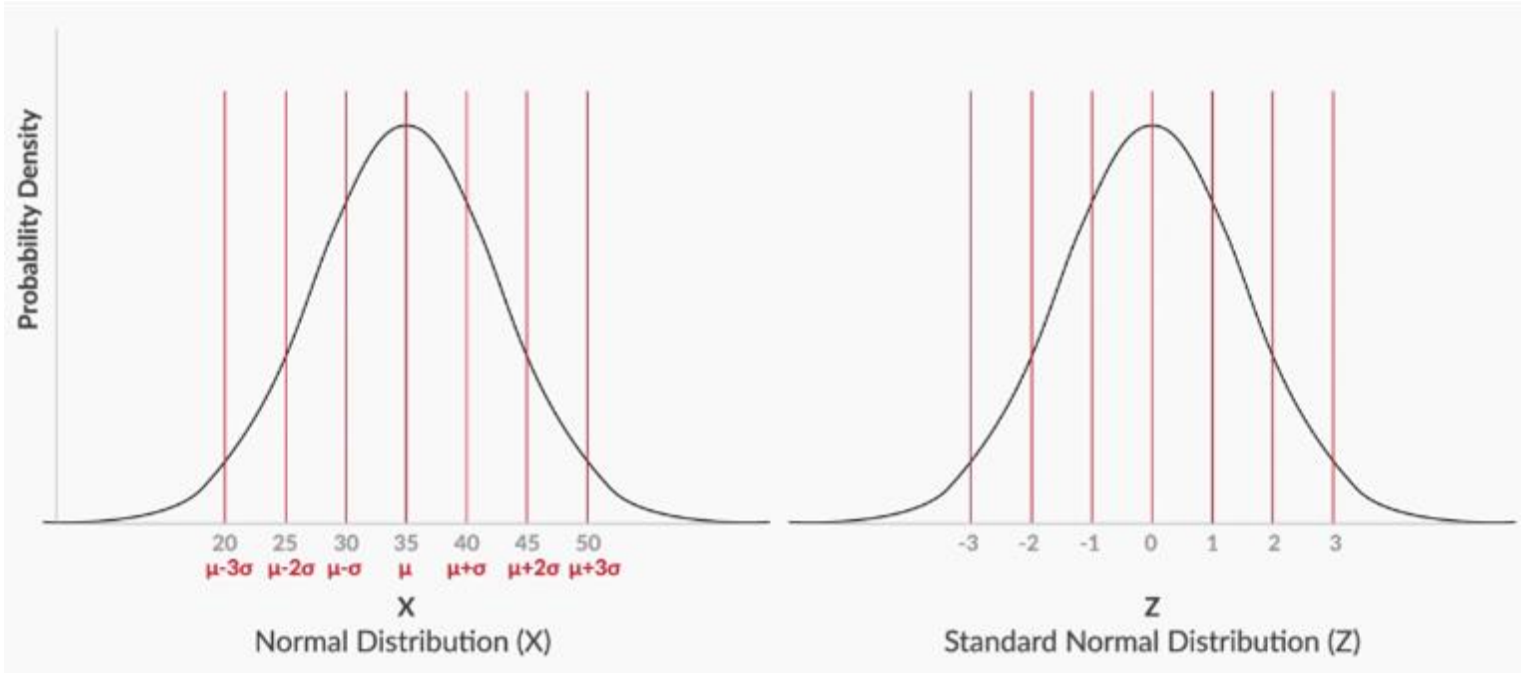


Calculating Probability with Z-score for Normal Distribution

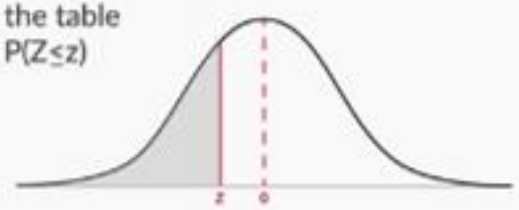
- If you want to find the probability, first step is to find out how far the value of X is from μ (in terms of "number of SD")
- The number is named as "z-score", "standard score" or "z value"

$$Z = \frac{X - \mu}{\sigma}$$

Basically, it tells you **how many standard deviations away from the mean** your random variable is.

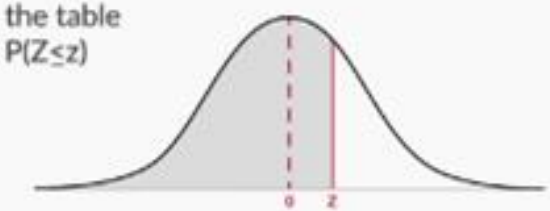


Number in the table represents P(Z≤z)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Number in the table represents P(Z≤z)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Example – Height Distribution in a Class

Imagine you are doing a study on the height of a class of university students.

Assuming that the height is following a normal distribution. What is the % of students who are shorter than 1.5m?

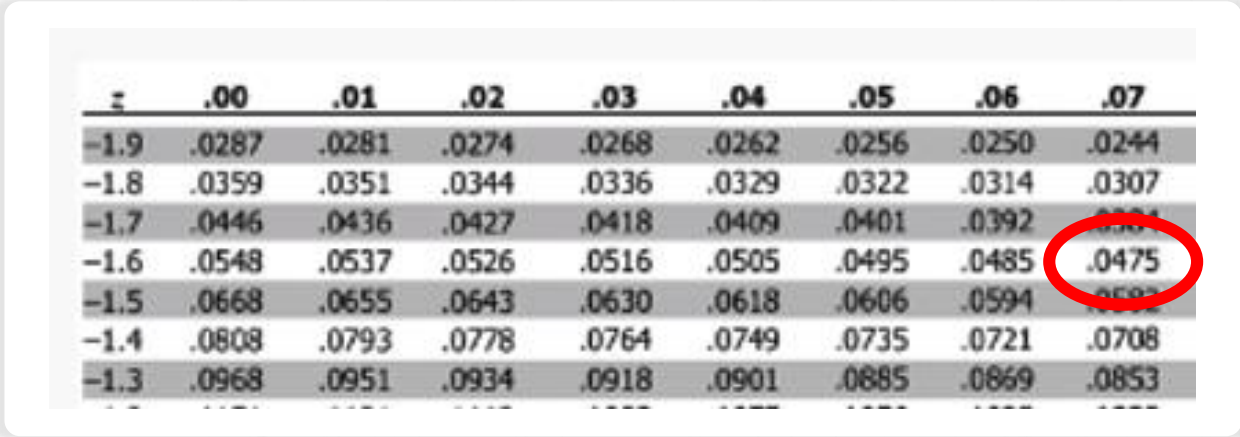
Average Height of the Class = 1.6m
Standard Deviation of Height of the Class = 0.06m

Step 1: Calculate the Z-score of the target height

$$Z - score\ of\ 1.5m = \frac{X\ -\ \mu}{\sigma} = \frac{1.5\ -1.6}{0.06} = -1.67$$

Step 2: Look up for the probability from the Z-score table.

$$P(height < 1.5m) = 0.0475 = 4.75\%$$



z	.00	.01	.02	.03	.04	.05	.06	.07
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853

Covariance & Correlation

Covariance

Covariance is a statistical term that refers to a relationship between two random variables on how one variable changes would impact the other one.

The covariance value can range from $-\infty$ to $+\infty$.

Positive covariance denotes a direct relationship and is represented by a positive number. If one variable is larger, then other one is also larger.

A negative covariance indicates an inverse relationship between the two variables. If one variable is larger, then other one is smaller.

The value does not represent the magnitude of the relationship. Only the direction matters.

Covariance

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

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.

Covariance Example

Stock Market Movement of 2 relevant stocks in the same industry

Day	Stock 1	Stock 2
1	+3%	+5%
2	+2%	+4.5%
3	+6%	+9%
4	+1%	+1%
5	+0%	+1%

Correlation

Correlation is similar to Covariance, also for measuring how two variable are related or moving together

Ranges from +1 to -1

Positive correlation: when one variable is bigger, the other one is also bigger

Negative correlation: when one variable is bigger, the other one is smaller

Both direction and magnitude are useful. The bigger the number, the stronger the relationship

Calculation of Correlation

$$\rho_{xy} = Correlation\,(x, y) = \frac{cov(x, y)}{\sqrt{var(x)}\ \sqrt{var(y)}}$$

cov(x, y): covariance of x and y

var(x): variance of x

var(y): variance of y

Covariance Example

Stock Market Movement of 2 relevant stocks in the same industry

Day	Stock 1	Stock 2
1	+3%	+5%
2	+2%	+4.5%
3	+6%	+9%
4	+1%	+1%
5	+0%	+1%

From previous section:

Cov(x, y) = 7.45

$$\rho_{xy} = Correlation\,(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\ \sqrt{var(y)}.}$$

Interpretation of correlation

Stock Market Movement of 2 relevant stocks in the same industry

Correlation Value (Absolute Value)	Meaning
0 - 0.3	Weak linear relationship
0.3 - 0.7	Moderate linear relationship
0.7 – 1	Strong linear relationship

Note: Correlation can only represent linear relationships, not other more complex relationships

Applications of Correlation

Explore what are the variables that are correlated among a large number of variables.

Example:
You have 100 variables about a person, which one is related to their online purchase habit?

User ID	Age	Gender	Phone Brand	Average Daily Online Time	Monthly Income	....	Number of Online Purchase
1	23	F	Apple	8	2000		20
2	46	M	Samsung	5	5000		2
3	55	F	Nokia	4	4000		10
4	18	M	Xiaomi	7	0		5
...							

Covariance vs Correlation

	Covariance	Correlation
Meaning	Measure how two variables are moving or correlated to each other	
Values	-∞ and +∞	-1 and +1
Direction	Positive/negative represent how two variables are moving together	
Magnitude/value	Not useful, not indicating the magnitude of the relationship	The value represents how strong the relationship is, and
Can compare for different group of variables based on the value (e.g. whether A/B have a strong relationship compare to A/B)	No	Yes

Hypothesis Testing

What is hypothesis testing ?

Hypothesis

- An assumption that we make about the population parameter.
  - Example 1: Average student in class is 40
  - Example 2: boys are taller than girls.

Hypothesis Testing

- A statistical method that is used in making statistical decisions using experimental data.
- All those example we assume need some statistic way to prove those.
- We need some mathematical conclusion whatever we are assuming is true.

Why do we use it ?

- An essential procedure in statistics.
- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- When we say that a finding is statistically significant, it's thanks to a hypothesis test.

Important parameter of hypothesis testing

Null hypothesis

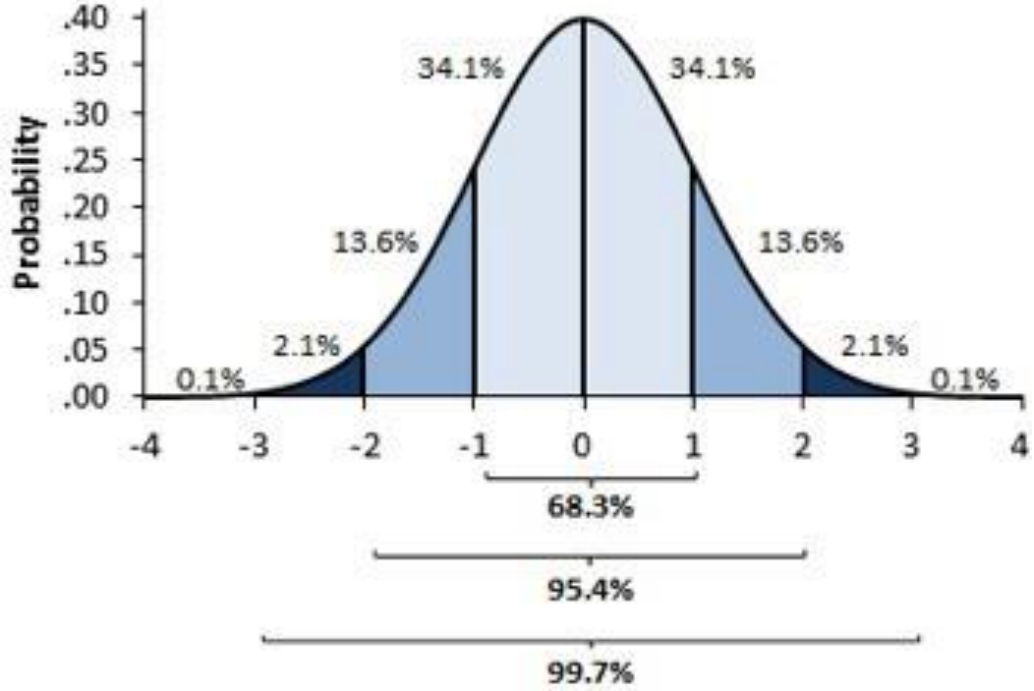
A general statement or default position that there is no relationship between two measured phenomena, or no association among groups In other words it is a basic assumption or made based on domain or problem knowledge.

Example : a company production is = 50 unit/per day etc.

Alternative hypothesis

The hypothesis used in hypothesis testing that is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect (with some amount of chance variation superposed)

Example : a company production is !=50 unit/per day etc.



Constructing Hypothesis

Null Hypothesis

- Greater than equal
- Less than equal

Alternate Hypothesis

- Not equal
- Less than
- Greater than

Example

Amazon's revenue last year was at least $14 billion

- Ho: Amazon's revenue >= 14 billion
- $H1$: Amazon's revenue < 14 billion

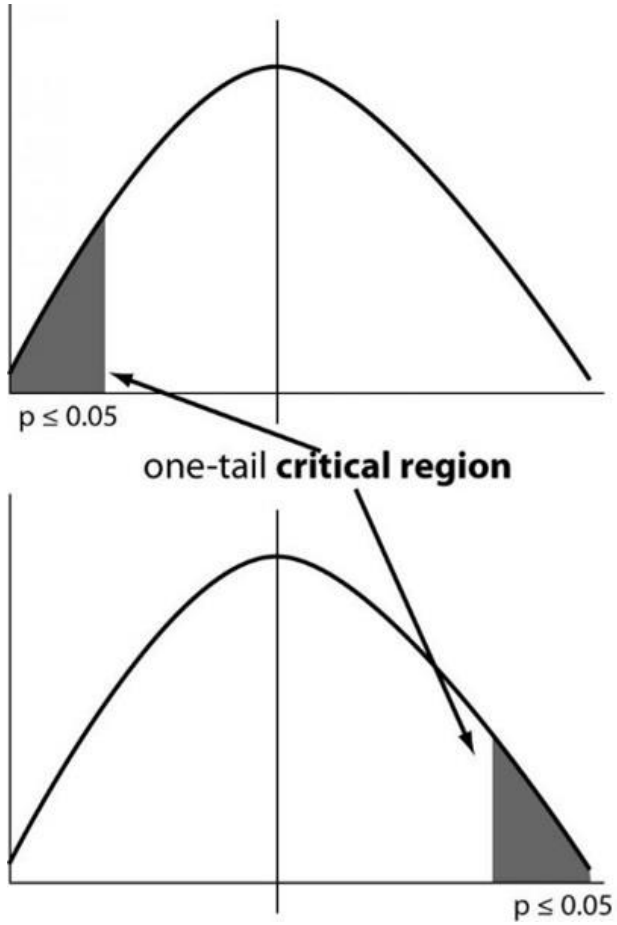
- Ho: Revenue >= 14
- $H1$: Revenue < 14

One tailed Test

A test of a statistical hypothesis , where the region of rejection is on only one side of the sampling distribution , is called a one-tailed test.

Example:

A student has ≥ 800 marks or math score ≥ 80%



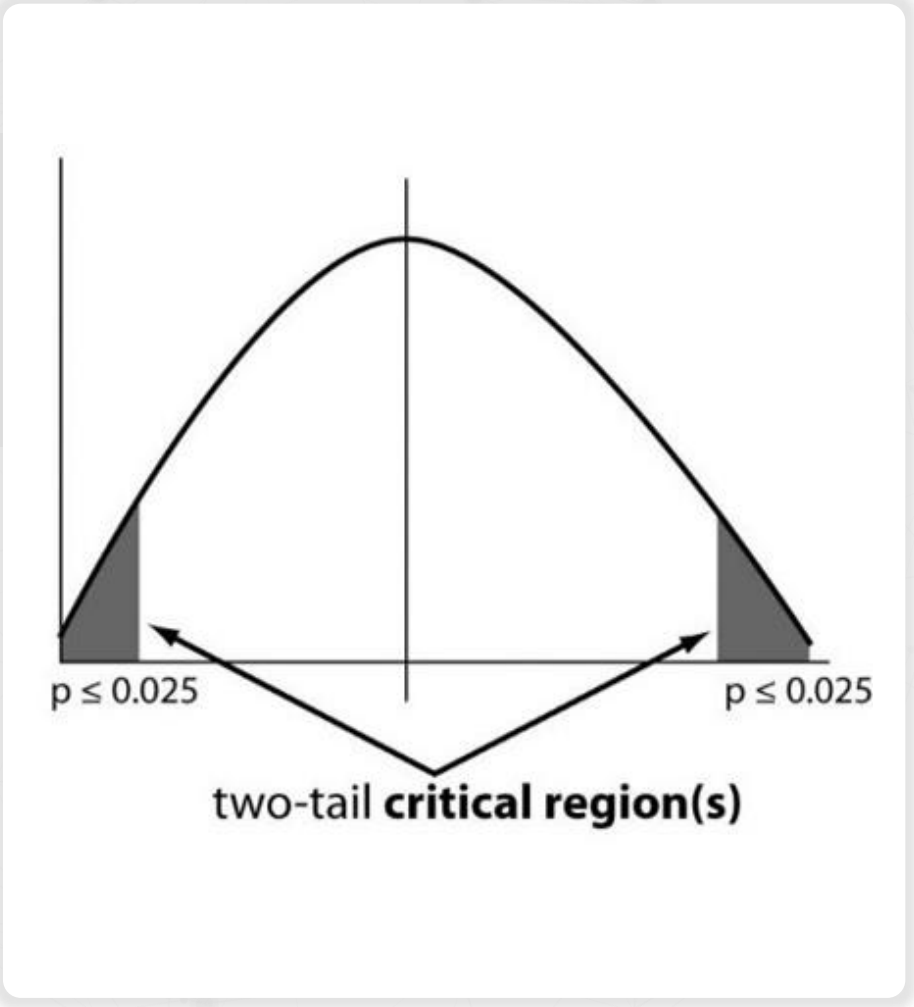
Two-tailed test

A two-tailed test is a statistical test in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

Example:

A student != 800 marks or mark score != 80%



P-value

The P value:

Or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true — the definition of 'extreme' depends on how the hypothesis is being tested.

Example

You have a coin and you don't know whether that is fair or tricky so let's decide null and alternate hypothesis

- H_0: a coin is a fair coin.
- H_1: a coin is a tricky coin. and alpha = 5% or 0.05

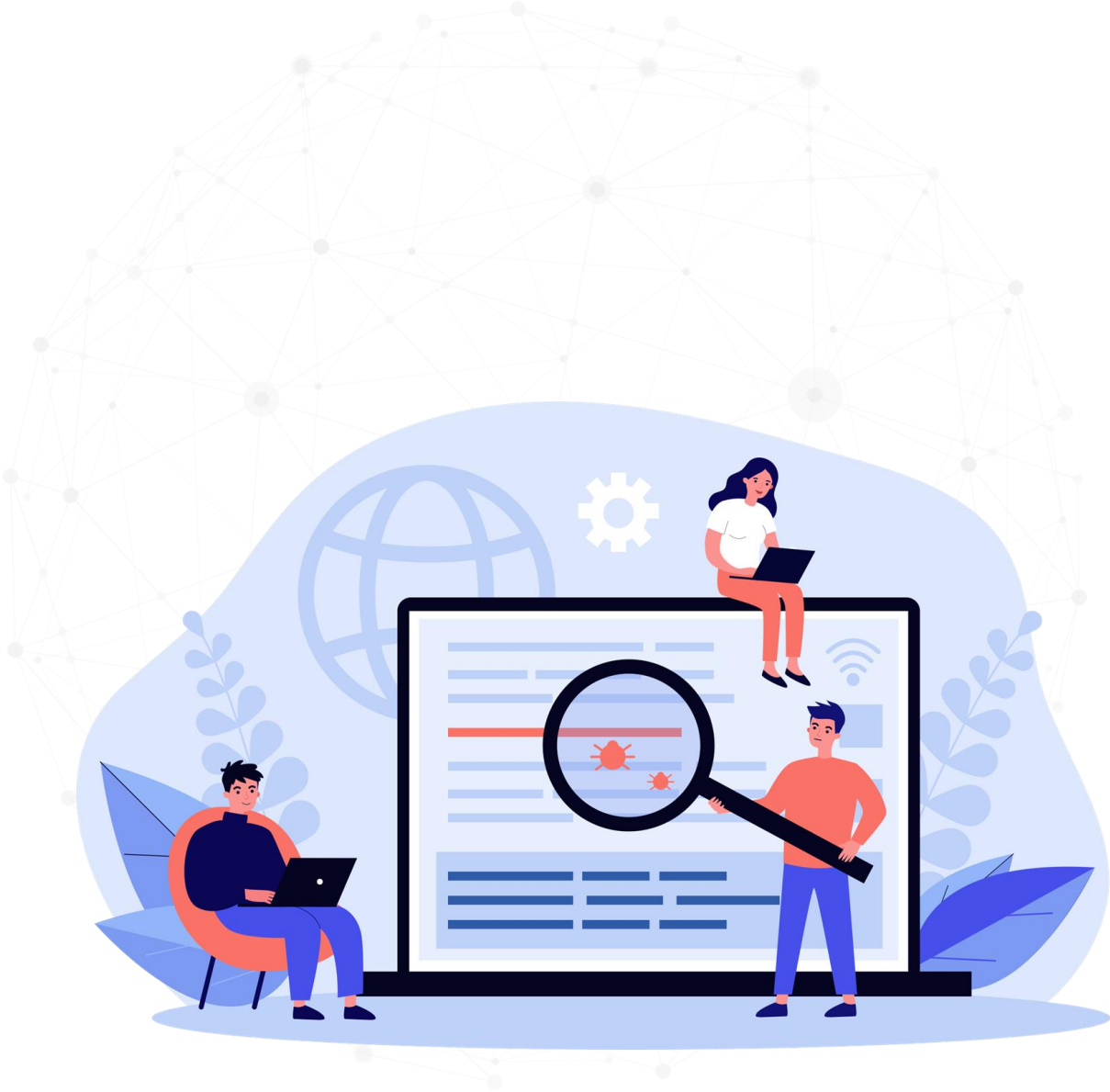
Now let's toss the coin and calculate p- value (probability value).

- Toss a coin 1st time and result is tail - P-value = 50% (as head and tail have equal probability)
- Toss a coin 2nd time and result is tail, now p-value = 50/2 = 25% and similarly we,
- Toss 6 consecutive time and got result as P-value = 1.5% but we set our significance level as 95% means 5% error rate we allow and here we see we are beyond that level

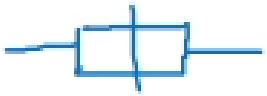
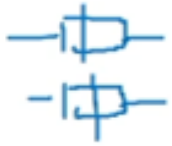
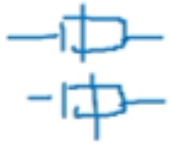
  i.e. our null hypothesis does not hold good so we need to propose that this coin is a tricky coin.

Types of Tests

1. T-test
2. Z-test
3. ANOVA Test
4. Chi-square Test
5. Correlation Test



Types of Tests for Different Data Types

Type of Data	What we see	Type of Test
1 Categorical		Z-Test or T-Test (depends if we have population variance or sample size >30)
1 Numerical		Z-Test or T-Test (depends if we have population variance or sample size >30)
1 Numerical + 1 Categorical (Less than 2 categories)		T-test or ANOVA
1 Numerical + 1 Categorical (More than 2 categories)		ANOVA
2 Categorical		Chi-square Test
2 Numerical		Correlation Test

Thank you