Advanced Analytics using Statistics PG-DBDA March 24

Session 1 & 2

Analytics is the systematic exploration and analysis of data to uncover meaningful patterns, insights, and trends that can inform decision-making and drive improvements in various aspects of business, science, and other domains. It involves the use of statistical, mathematical, and computational techniques to extract actionable insights from data.

Understanding Data: Analytics begins with understanding the data available. This includes both structured data (such as databases and spreadsheets) and unstructured data (such as text documents, images, and videos). The data can come from various sources, including business transactions, customer interactions, sensors, social media, and more.

Data Preparation: Before analysis can begin, the raw data often needs to be cleaned, transformed, and formatted to make it suitable for analysis. This process involves tasks such as removing duplicates, handling missing values, standardizing formats, and integrating data from different sources.

Descriptive Analytics: Descriptive analytics focuses on summarizing and describing the characteristics of the data. This may involve generating summary statistics, visualizing data through charts and graphs, and exploring relationships between variables. Descriptive analytics helps stakeholders understand what has happened in the past and provides context for further analysis.

Diagnostic Analytics: Diagnostic analytics aims to understand why certain events occurred by identifying patterns and correlations in the data. It involves digging deeper into the data to uncover the root causes of observed phenomena or trends. Diagnostic analytics often involves hypothesis testing, correlation analysis, and causal inference techniques.

Predictive Analytics: Predictive analytics leverages historical data to forecast future outcomes or trends. This involves building statistical or machine learning models that can make predictions based on patterns observed in the data. Predictive analytics can be used for various purposes, such as sales forecasting, customer churn prediction, risk assessment, and demand forecasting.

Prescriptive Analytics: Prescriptive analytics goes beyond prediction to recommend actions or decisions that can optimize outcomes. This involves using optimization and simulation techniques to explore different scenarios and identify the best course of action given specific constraints and objectives. Prescriptive analytics helps organizations make data-driven decisions to improve efficiency, minimize risks, and maximize outcomes.

Continuous Improvement: Analytics is an iterative process that requires continuous improvement and refinement. As new data becomes available and business conditions change, analytics models and strategies need to be updated and adapted accordingly. Organizations should establish feedback loops to incorporate insights gained from analytics into decision-making processes and drive ongoing improvement.

Overall, analytics enables organizations to leverage data as a strategic asset to gain competitive advantage, improve operational efficiency, enhance customer experiences, and drive innovation. By harnessing the power of analytics, businesses and other entities can make smarter decisions and achieve better outcomes in an increasingly data-driven world.

Data analytics Life Cycle

The data analytics lifecycle is a structured approach to extracting insights and value from data. It typically consists of several interconnected stages that guide the process from defining the problem to implementing solutions. Here's a breakdown of the data analytics lifecycle:

Problem Definition:

- Identify the business problem or opportunity that analytics can address.
- Define clear objectives and key performance indicators (KPIs) to measure success.
- Ensure alignment with organizational goals and stakeholder needs.

Data Collection:

- Identify relevant data sources both internal and external to the organization.
- Gather data from databases, spreadsheets, files, APIs, sensors, social media, etc.
- Ensure data quality, completeness, and relevance for analysis.

Data Preparation:

- Cleanse the data by removing duplicates, correcting errors, and handling missing or inconsistent values.
- Transform the data into a suitable format for analysis (e.g., normalization, aggregation, or feature engineering).
- Integrate data from multiple sources if necessary.

Exploratory Data Analysis (EDA):

- Explore the dataset to understand its structure, distribution, and relationships.
- Visualize data using charts, graphs, and statistical summaries.
- Identify patterns, trends, outliers, and potential insights.

Feature Engineering:

- Select, create, or transform features that are relevant and predictive for the analysis.
- Apply techniques such as dimensionality reduction, encoding categorical variables, or deriving new features.

Modeling:

- Select appropriate analytical techniques or algorithms based on the problem and data characteristics.
- Split the data into training, validation, and testing sets.
- Train machine learning or statistical models using the training data.
- Tune hyperparameters and evaluate model performance using validation data.
- Validate the model's performance on unseen data using the testing set

Interpretation and Evaluation:

- Interpret the model results in the context of the problem and business objectives.
- Evaluate the model's performance using relevant metrics (e.g., accuracy, precision, recall, or AUC).
- Assess the impact of the analytics solution on the business problem and its alignment with KPIs.

Deployment:

- Deploy the analytics solution into production or operational systems.
- Integrate the model into decision-making processes or business workflows.
- Monitor the model's performance in real-world scenarios and collect feedback for continuous improvement.

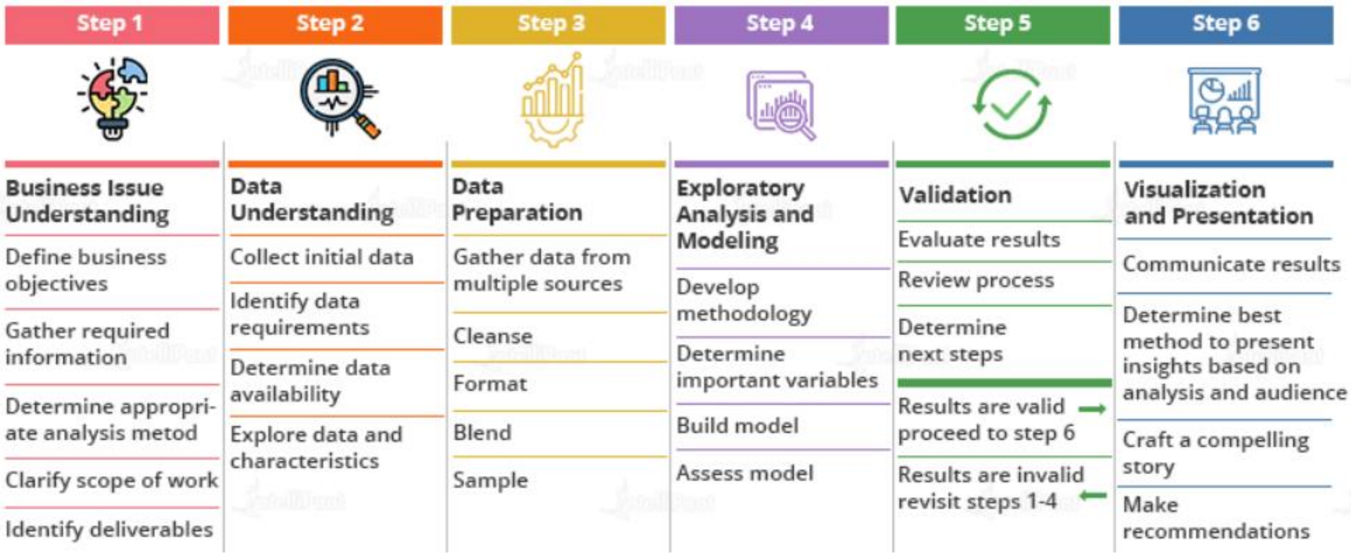
Monitoring and Maintenance:

- Establish monitoring mechanisms to track the performance and behavior of the deployed model.
- Monitor data quality, model drift, and other relevant metrics over time.
- Retrain or update the model periodically with new data to ensure relevance and accuracy.

Iterative Improvement:

- Continuously refine and improve the analytics solution based on feedback, changing business requirements, and new data.
- Iterate through the lifecycle stages as needed to address evolving challenges and opportunities.

By following the data analytics lifecycle, organizations can systematically leverage data to derive actionable insights, make informed decisions, and drive business value.



Step 1
Business Issue Understanding
Define business objectives
Gather required information
Determine appropri-ate analysis metod
Clarify scope of work
Identify deliverables
Step 2
Data Understanding
Collect initial data
Identify data requirements
Determine data availability
Explore data and characteristics
Step 3
Data Preparation
Gather data from multiple sources
Cleanse
Format
Blend
Sample
Step 4
Exploratory Analysis and Modeling
Develop methodology
Determine important variables
Build model
Assess model
Step 5
Validation
Evaluate results
Review process
Determine next steps
Results are valid proceed to step 6
Results are invalid revisit steps 1-4
Step 6
Visualization and Presentation
Communicate results
Determine best method to present insights based on analysis and audience
Craft a compelling story
Make recommendations

Session 3 & 4

Prerequisites:

Factorial Notation:

We define, $0! = 1$

For any positive integer n,

$n! = n\times(n-1)\times(n-2)\times\ldots\times 1$

e.g.

$1! = 1$

$2! = 2\times1$

$3! = 3\times2\times1$

$4! = 4\times3\times2\times1$ and so on.

Consider $6! = 6\times5\times4\times3\times2\times1$, which we can write as

$6! = 6\times\mathbf{(5\times4\times3\times2\times1)} = 6\times\mathbf{5!}$ $\therefore \mathbf{n! = n\times[(n-1)!]}$

$6! = 6\times5\times\mathbf{(4\times3\times2\times1)} = 6\times5\times\mathbf{4!}$

$\therefore \mathbf{n! = n\times(n-1)\times[(n-2)!]}$ and so on.

Permutation:

Consider selection of 'r' objects out of n ($r \le n$). If the **order** in which the objects are selected **is important** then such a selection is called as a **Permutation.**

The number of such permutations is denoted by ${}^{n}P_{r}$ and

$$\text{nPr} = \frac{n!}{(n-r)!}$$

e.g.

$$\text{nP0} = \frac{n!}{(n-0)!} = 1$$

$$\text{nP1} = \frac{n!}{(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$$

$$\text{nPn} = \frac{n!}{(n-n)!} = \text{n!}$$

Combination:

Consider selection of 'r' objects out of n ($r \le n$). If the **order** in which the objects are selected **is not important** then such a selection is called as a **Combination.**

The number of such combinations is denoted by $^{n}C_{r}$ and

$$\text{nCr} = \frac{n!}{r!\,(n-r)!}$$

e.g.

$$\text{nC0} = \frac{n!}{0!\,(n-0)!} = \frac{n!}{n!} = 1$$

$$\text{nC1} = \frac{n!}{1!\,(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$$

$$\text{nCn} = \frac{n!}{1!\,(n-n)!} = \frac{n!}{n!\,0!} = 1$$

$$10\text{C}3 = \frac{10!}{3!\,(10-3)!} = \frac{10!}{3!\,7!}$$

$$10\text{C}7 = \frac{10!}{7!\,(10-7)!} = \frac{10!}{7!\,3!}$$

$$\therefore {}^{10}C_3 = {}^{10}C_7 \text{ i.e. } {}^{10}C_3 = {}^{10}C_{10-3}$$

$$\boxed{\textbf{In general, } {}^{n}C_{r} = {}^{n}C_{n-r}}$$

$${}^{10}C_3 = \frac{10!}{3!7!} = \frac{10*9*8*7!}{3!7!} = \frac{10*9*8}{3!}\ ; \quad {}^{12}C_4 = \_\_\_\_\_\_\_\_\_\_\_\_$$

$${}^{100}C_{97} = {}^{100}C_3 = \_\_\_\_\_\_\_\_\_\_\_\_$$

Basic Terms:

Random Experiment:

Consider an action which is repeated under essentially identical conditions. If it results in any one of the several possible outcomes, but it is not possible to predict which outcome will appear, then such an action is called as a Random Experiment

A random experiment is defined as an experiment whose outcome cannot be predicted with certainty

An activity that produces a result or an outcome is called an experiment. It is an element of uncertainty as to which one of these occurs when we perform an activity or experiment. Usually, we may get a different number of outcomes from an experiment. However, when an experiment satisfies the following two conditions, it is called a random experiment.

(i) It has more than one possible outcome.

(ii) It is not possible to predict the exact outcome in advance.

Outcome

A possible result of random experiment is called a possible outcome of the experiment.

Sample Space:

The set of all possible outcomes of a random experiment is called the sample space. The sample space is denoted by S or Greek letter omega (Ω). The number of elements in S is denoted by n(S). A possible outcome is also called a sample point since it is an element in the sample space.

All the elements of the sample space together are called as 'exhaustive cases'.

Event:

Any subset of the sample space is called as an 'Event' and is denoted by any capital letter like A, B, C or A_1, A_2, A_3,..

Favourable cases:

The cases which ensure the happening of an event A, are called as the cases favourable to the event A. The number of cases favourable to event A is denoted by n(A).

Types of Events

Elementary Event: An event consisting of a single outcome is called an elementary event.

Certain Event: The sample space is called the certain event if all possible outcomes are favourable outcomes. i.e. the event consists of the whole sample space.

Impossible Event: The empty set is called impossible event as no possible outcome is favorable

Union of Two Events

Let A and B be two events in the sample space S. The union of A and B is denoted by $A \cup B$ and is the set of all possible outcomes that belong to at least one of A and B.

Let S = Set of all positive integers not exceeding 50;

Event A = Set of elements of S that are divisible by 6;

Event B = Set of elements of S that are divisible by 9.

$A = \{6,12,18,24,30,36,42,48\}$

$B = \{9,18,27,36,45\}$

$\therefore A \cup B = \{6,9,12,18,24,27,30,36,42,45, 48\}$ is the set of elements of S that are divisible by 6 or 9.

Exhaustive Events

Two events A and B in the sample space S are said to be exhaustive if $A \cup B = S$

Intersection of Two Events

Let A and B be two events in the sample space S.

The intersection of A and B is the event consisting of outcomes that belong to both the events A and B.

Let S = Set of all positive integers not exceeding 50,

Event A = Set of elements of S that are divisible by 3,

Event B = Set of elements of S that are divisible by 5.

Then $A = \{3,6,9,12,15,18,21,24,27,30,33, 36,39,42,45,48\}$,

$B = \{5,10,15,20,25,30,35,40,45,50\}$

$\therefore A \cap B = \{15,30,45\}$ is the set of elements of S that are divisible by both 3 and 5.

Mutually Exclusive Events

Event A and B in the sample space S are said to be mutually exclusive if they have no outcomes in common. ($A \cap B = \phi$). In other words, the intersection of mutually exclusive events is empty. Mutually exclusive events are also called disjoint events.

If two events A and B are mutually exclusive and exhaustive, then they are called **Complementary events**.

Equally Likely Cases:

Cases are said to be equally likely if they all have the same chance of occurrence i.e. no case is preferred to any other case.

Probability Introduction

Chance is the occurrence of events in the absence of any obvious intention or cause. It is, simply, the possibility of something happening. When the chance is defined in Mathematics, it is called probability.

Probability is the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible.
Mathematically, the probability of an event occurring is equal to the ratio of a number of cases favourable to a particular event to the number of all possible cases.

The theoretical probability of an event is denoted as P(E).

$$P(E) = \frac{\text{Number of Outcomes Favourable to E}}{\text{Number of all Possible Outcomes of the Experiment}}$$

Importance of Probability

The concept of probability is of great importance in everyday life. Statistical analysis is based on this valuable concept. Infact the role played by probability in modern science is that of a substitute for certainty.

The following discussion explains it further:

i. The probability theory is very much helpful for making prediction. Estimates and predictions form an important part of research investigation. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.

ii. It has also immense importance in decision making.

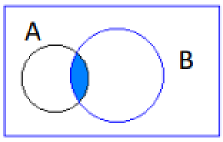
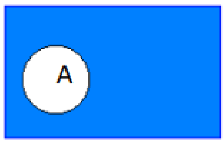
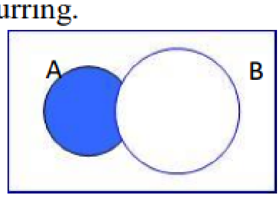
iii. It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.

iv. It is one of the inseparable tools for all types of formal studies that involve uncertainty.

v. The concept of probability is not only applied in business and commercial lines, rather than it is also applied to all scientific investigation and everyday life.

vi. Before knowing statistical decision procedures one must have to know about the theory of probability.

vii. The characteristics of the Normal Probability. Curve is based upon the theory of probability.

S.NO	Operator	Symbol	Example	Meaning
1	Union	$\cup$	$A \cup B$	The event of either A or B occurring.
2	Finite union	$\bigcup_{i=1}^{n} A_i$	$\bigcup_{i=1}^{3} A_i$	The event of any one of the events A_1, A_2 and A_3 occurring.
3	Countable union	$\bigcup_{i=1}^{\infty} A_i$	$\bigcup_{i=1}^{\infty} A_i$	The event of any one of the events A_1,A_2 ….. occurring.
4	Intersection	$\cap$	$A \cap B$	The event of both A and B occurring. 
5	Finite intersection	$\bigcap_{i=1}^{n} A_i$	$\bigcap_{i=1}^{3} A_i$	The event of all the events A_1,A_2 and A_3 occurring.
6	Countable intersection	$\bigcap_{i=1}^{\infty} A_i$	$\bigcap_{i=1}^{\infty} A_i$	The event of all the events A_1,A_2 ….. occurring.
7	Complementation	c or –	A^c or $\bar{A}$	The event of A not occurring. 
8	Subtraction	-	A-B	The event of A occurring and B not occurring. 

Operation	Interpretation
A', A or A^c	Not A.
$A \cup B$	At least, one of A and B
$A \cap B$	Both A and B
$(A' \cap B) \cup (A \cap B')$	Exactly one of A and B
$(A' \cap B') = (A \cup B)'$	Neither A nor B

Elementary Properties of Probability:

1) A' is complement of A and therefore $P(A') = 1 - P(A)$

2) For any event A in S, $0 \le P(A) \le 1$

3) For the impossible event φ, $P(\varphi) = 0$

4) For the certain event S, $P(S) = 1$

5) If A1 and A2 two mutually exclusive events then $P(A1 \cup A2) = P(A1) + P(A2)$

6) If $A \subseteq B$, then $P(A) \le P(B)$ and $P(A' \cap B) = P(B) - P(A)$

7) Addition theorem: For any two events A and B of a sample space S,

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

8) For any two events A and B, $P(A \cap B') = P(A) - P(A \cap B)$

9) For any three events A, B and C of a sample space S,

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

10) If $A_1, A_2, \ldots\ldots, A_m$ are mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup, \ldots\ldots \cup A_m = P(A_1) + P(A_2) + \ldots. + P(A_m)$$

Ex-1: If a die is rolled find the probability that number on the uppermost face of the die is

a) an odd no.
b) a prime no.
c) greater than 2.

Ex-2: If three coins are tossed simultaneously, find the probability of getting

a) exactly one head
b) at least one head
c) no head.

Ex-3: Find the probability that a leap year selected at random contains 53 Sundays.

Ex-4: In a housing society, half of the families have a single child per family, while the remaining half have two children per family. If a child is picked at random, find the probability that the child has a sibling.

Ex-5: A box contains 6 white and 4 black balls. 2 balls are selected at random and the colour is noted. Find the probability that

a) Both balls are white
b) Both balls are black.
c) Balls are of different colours.

Ex-6: If all the letters of the word EAR are arranged at random. Find the probability that the word begins and ends with a vowel.

Ex-7: If all the letters of the word EYE are arranged at random find the probability that the word begins and ends with vowels.

Ex-8: If all the letters of word EQUATION are arranged at random, find the probability that the word begins and ends with a vowel.

Ex-9: 7 boys and 3 girls are arranged in a row. Find the probability that there is at least one boy between 2 girls.

Ex-10: 6 books on accountancy, 5 books of economics and 4 books of mathematics are to be arranged in the shelf find the probability that all the books of one subject are together.

Complement of an event: The complement of an event A is denoted by $\bar{A}$ or A′ or A^C and it contains all the elements of the sample space which do not belong to A.
e.g.random experiment: an unbiased die is rolled.
S = {1, 2, 3, 4, 5, 6}
(i) Let A: number on the die is a perfect square
$\therefore$ A = {1, 4} $\therefore$ $\bar{A}$ = {2, 3, 5, 6}
(ii) Let B: number on the die is a prime number
$\therefore$ B = {2, 3, 5} $\therefore$ $\bar{A}$ = {1, 4, 6}

Note:
$P(A) + P(\bar{A}) = 1$
i.e. $P(A) = 1 - P(\bar{A})$

Ex-11: If 3 dice are tossed simultaneously, find the probability that the sum of the 3 numbers is less than 17.

Addition Theorem of Probability:

Result: If A and B are any two events then
$P(A\cup B) = P(A) + P(B) - P(A\cap B)$

Note:
1) $A\cup B$: either A or B or both;
$A\cup B$: at least one of A & B

2) If A & B are mutually exclusive, $P(A\cap B) = 0$
$\therefore P(A\cup B) = P(A) + P(B)$

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

Ex-13: The probability that a particular film gets award for best direction is 0.7. The probability that it gets award for best acting is 0.4. The probability that the film gets award for both is 0.2. Find the probability that the film gets
a) at least one award
b) no award

Ex-14: Two cards are selected at random from a pack of 52 cards, find the probability that two cards are
a) Red or face cards
b) Aces or jacks.

Conditional Probability

Let S be a sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of event A under the condition that event B has already occurred and $P(B) \neq 0$ is called conditional probability of event A given B and is denoted by P(A/B).

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$$

Multiplication theorem

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of both the events is denoted by $P(A \cap B)$

and is given by $P(A \cap B) = P(A).P(B/A)$

Independent Events

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S. If the occurrence of either event, does not affect the probability of the occurrence of the other event, then the two events A and B are said to be independent.

Thus, if A and B are independent events then,

$P(A/B) = P(A/B') = P(A)$ and $P(B/A) = P(B/A') = P(B)$

If A and B are independent events then $P(A \cap B) = P(A).P(B)$

$(P(A \cap B) = P(A).(B/A) = P(A).P(B) \therefore P(A \cap B) = P(A).P(B))$

If A and B are independent events then

a) A and B' are also independent event

b) A' and B' are also independent event

Ex-15: 2 shooters are firing at target. The probability that they hit the target are 1/3 and 1/2 respectively. If they fire independently find the probability that

a) both hit the target.
b) Nobody hits the target.
c) At least one hits the target.
d) Exactly one hits the target.

Ex-17: Suppose A & B are two independent events such that
$P(A) = 0.4$, $P(A \cup B') = 0.7$. Find $P(A \cup B)$.

Ex-18: Three vendors were asked to supply a component. The respective probabilities that the component supplied by them is 'good' are 0.8, 0.7 and 0.5. Each vendor supplies only one component. Find the probability that at least one component is 'good'.

Ex-19: The chance of a student passing a test is 20%. The chance of student passing the test and getting above 90% marks is 5%. Given that a student passes the test, find the probability that the student gets above 90% marks.

Ex-20: A box contains 6 white and 4 black balls. One ball is selected at random and its colour is noted. The ball is replaced and two balls of the opposite colour are added and then second ball is selected at random find the probability that both balls are white.

Ex-21: A shop has equal number of LED bulbs of two different types. The probability that the life of an LED bulb is more than 100 hours given that it is of type-1 is 0.7 and given that it is of type-2 is 0.4. If an LED bulb is selected at random, find the probability that the life of the bulb is more than 100 hours.

Bayes Theorem

Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence

- Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Bayes' Theorem was named after 18th-century mathematician Thomas Bayes.
- It is often employed in finance in calculating or updating risk evaluation.
- The theorem has become a useful element in the implementation of machine learning.
- The theorem was unused for two centuries because of the high volume of calculation capacity required to execute its transactions.

Applications of Bayes' Theorem are widespread and not limited to the financial realm. For example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Bayes' Theorem formula

$$\underbrace{P(A|B)}_{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{likelihood}} * \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{evidence}}}$$

<u>Ex-22:</u> In a bolt factory, three machine P, Q and R produce 25%, 35% and 40% of the total output respectively. It is found that in their production, respectively 5%, 4% and 2% are defective bolts. If a bolt is selected at random and found defective, find the probability that it is produced by machine Q.

<u>Ex-23:</u> A certain test for a particular cancer is known to be 95% accurate. A person submits to the test and the results are positive. Suppose that the person comes from a population of 1,00,000 where 2,000 people suffer from that disease. What can we conclude about the probability that the person under test has that particular disease?

ODDS (Ratio of two complementary probabilities):

Let n be number of distinct sample points in the sample space S. Out of n sample points, m sample points are favourable for the occurrence of event A. Therefore remaining (n-m) sample points are favourable for the occurrence of its complementary event A'.

$\therefore P(A) = \frac{m}{n}$ and $P(A') = \frac{n-m}{n}$

Ratio of number of favourable cases to number of unfavourable cases is called as odds in favour of event A which is given by $\frac{m}{n-m}$ i.e. P(A):P(A').

Ratio of number of unfavourable cases to number of favourable cases is called as odds against event A which is given by $\frac{n-m}{n}$ i.e. P(A'):P(A

Session 5 & 6

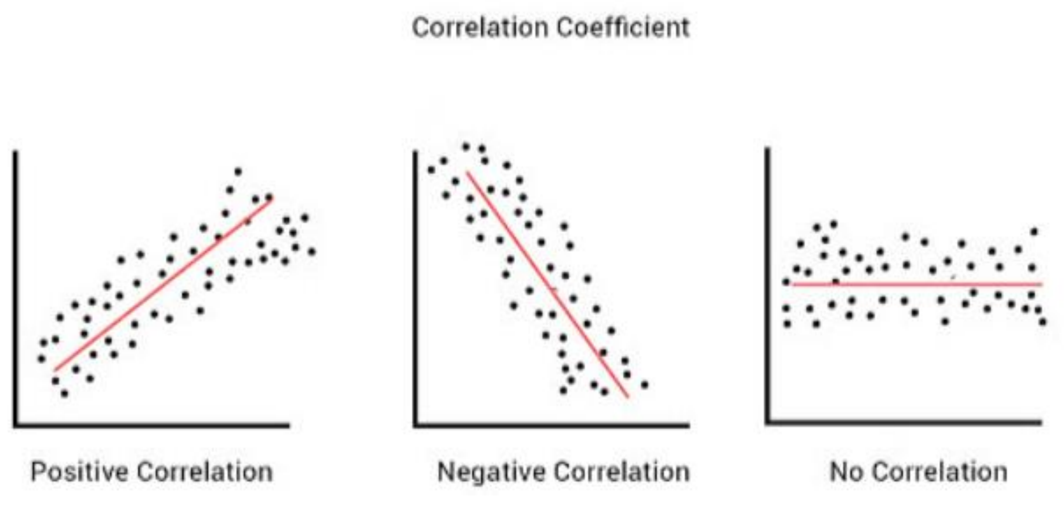
What Is Correlation?

Correlation refers to the statistical relationship between the two entities. It measures the extent to which two variables are linearly related. For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people.

You can apply correlation to a variety of data sets. In some cases, you may be able to predict how things will relate, while in others, the relation will come as a complete surprise. It's important to remember that just because something is correlated doesn't mean it's causal.

There are three types of correlation:

- Positive Correlation: A positive correlation means that this linear relationship is positive, and the two variables increase or decrease in the same direction.
- Negative Correlation: A negative correlation is just the opposite. The relationship line has a negative slope, and the variables change in opposite directions, i.e., one variable decreases while the other increases.
- No Correlation: No correlation simply means that the variables behave very differently and thus, have no linear relationship



Scatter diagram:

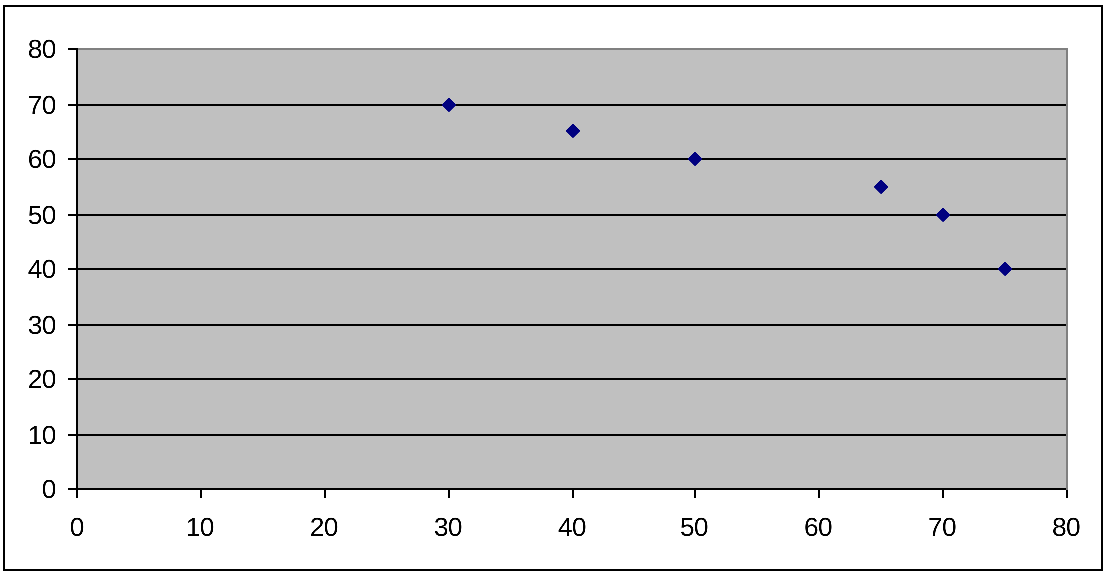
We collect a data of pairs of values of the two variables. Generally these variables are denoted by x & y. These values are considered as x & y co-ordinates respectively and plotted as points on a graph. Such diagrammatic representation of a bivariate data is called as a scatter diagram. From the scatter diagram a rough idea of the nature of relationship between the two variables can be drawn as follows.

Ex: Draw a scatter diagram for the following data and give your comments.

x	30	40	50	65	70	75
y	70	65	60	55	50	40

Answer: Scatter Diagram:

Comment: There is negative correlation between x & y



Ex: Draw a scatter diagram for the following data and comment.

Demand	15	20	18	22	25	30
Price	32	19	25	15	12	10

Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Note:

1) r lies between -1 & 1 i.e. $-1 \le r \le 1$
2) If $r = 1$, there is perfect positive correlation
3) If $0 < r < 1$, there is positive correlation
4) If $r = -1$, there is perfect negative correlation
5) If $-1 < r < 0$, there is negative correlation
6) If $r = 0$, there is no correlation
7) Correlation Coefficient is independent of change of origin & change of scale.

Ex: Calculate correlation coefficient for the following data. Comment on your findings.

Marks in Statistics	53	59	72	43	93	35	55	70
Marks in Economics	35	49	63	36	75	28	38	76

Ex: Calculate Karl Pearson's Coefficient of correlation for the following data.

X	17	8	12	13	10	12
Y	13	7	10	11	8	11

Ex: Find the Karl Pearson's correlation coefficient for the following data.

x	10	14	12	18	20	16
y	20	30	20	35	25	20

Spearman's Rank Correlation Coefficient:

In this method, ranks are assigned to the data. The ranks are given to the x-series & y-series separately. The highest observation is given rank '1', the next highest observation is given rank '2' and so on. Suppose, R_1 & R_2 are the ranks of the x & y respectively and $d = R_1 - R_2$ then

$$r = 1 - \left\{ \frac{6 \sum d^2}{n(n^2 - 1)} \right\}$$

where n = number of pairs of observations

Ex: Calculate the Spearman's rank correlation coefficient for the following data.

x	15	12	16	13	17	14	18	11
y	17	14	20	25	23	24	22	21

<u>Ex:</u> Calculate the Spearman's rank correlation coefficient for the following data.

x	50	63	40	70	45	65	38	53	52
y	48	30	35	60	55	33	25	54	50

"Causation is not correlation" is a fundamental concept in statistics and scientific research. It essentially means that just because two variables are correlated (meaning they tend to vary together), it doesn't necessarily mean that one causes the other to happen.

Here's an example to illustrate this:

Let's say we observe a strong positive correlation between ice cream sales and the number of drownings at the beach. During the summer months, both ice cream sales and drownings tend to increase. However, it would be incorrect to conclude that eating ice cream causes people to drown or vice versa.

There could be a third variable at play here, such as temperature. Warmer temperatures in the summer lead to increased ice cream consumption as well as more people going to the beach and swimming, which in turn increases the risk of drownings. So, in this example, temperature is the common cause behind both variables—ice cream sales and drownings—rather than one causing the other directly.

To establish causation, researchers often conduct controlled experiments or use sophisticated statistical methods to account for potential confounding variables. These methods help them determine if changes in one variable directly lead to changes in another variable, thus establishing a cause-and-effect relationship.

Covariance Meaning

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

Positive Covariance

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

Negative Covariance

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

Covariance Formula

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by Cov(X,Y). The formula is given below for both population covariance and sample covariance.

$$\mathrm{Cov}(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If cov(X, Y) is zero, then we can say that there is no relation between two variables.

Regression

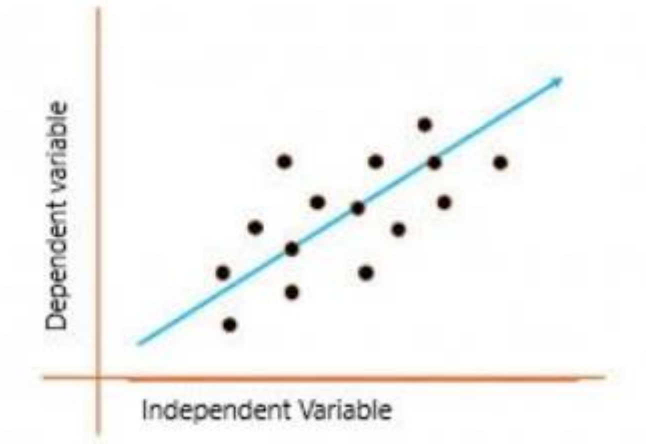
What is Linear Regression?

“Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values. Applied in various domains like economics and finance, this method analyzes and forecasts data trends. It can extend to multiple linear regression involving several independent variables and logistic regression, suitable for binary classification problems

Simple Linear Regression

In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.

Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable **X**(independent variable), such linear regression is ***simple linear regression***.



The graph above presents the linear relationship between the output(y) and predictor(X) variables. The blue line is referred to as the *best-fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

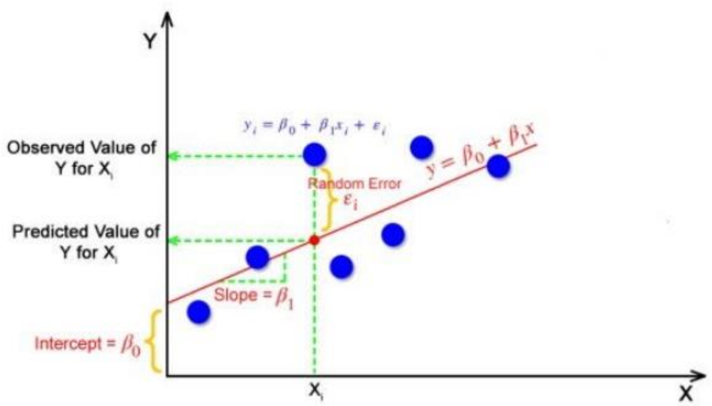
Simple Regression Calculation

To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$Y_i = \beta_0 + \beta_1 X_i$

where Y_i = Dependent variable, β_0 = constant/Intercept, β_1 = Slope/Intercept, X_i = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line $Y = B_0 + B_1 X$.



But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B_0 and B_1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

Random Error(Residuals)

In regression, the difference between the observed value of the dependent variable($\mathbf{y_i}$) and the predicted value(**predicted**) is called the residuals.

$$\varepsilon_i = y_{predicted} - y_i$$

where $y_{predicted} = B_0 + B_1 X_i$

Outliers

What is an outlier?

An outlier is an observation "that appears to deviate markedly from other members of the sample in which it occurs"

What causes outliers?

- Human errors, e.g. data entry errors
- Instrument errors, e.g. measurement errors
- Data processing errors, e.g. data manipulation
- Sampling errors, e.g. extracting data from wrong sources
- Not an error, the value is extreme, just a 'novelty' in the data

A dilemma

- Outliers can be genuine values
- The trade-off is between the loss of accuracy if we throw away "good" observations, and the bias of our estimates if we keep "bad" ones
- The challenge is twofold:

1. to figure out whether an extreme value is good (genuine) or bad (error)
2. to assess its impact on the statistics of interest

Outlier treatment is the process of identifying and handling outliers in a dataset. Outliers are defined as observations that fall outside of the general pattern of the data and can have a significant impact on the analysis and modeling of the data.

There are several methods for identifying and treating outliers, including:

Z-score method: This method calculates the standard deviation and mean of the data, and any observation that falls more than 3 standard deviations away from the mean is considered an outlier.

Interquartile range method: This method calculates the interquartile range (IQR) of the data, and any observation that falls outside of the lower and upper limits of the box plot, which are defined as Q1–1.5 * IQR and Q3 + 1.5 * IQR, respectively, is considered an outlier.

Clustering methods: Clustering methods such as DBScan and KMeans can also be used to identify outliers by grouping similar data points together and identifying any data points that do not belong to any cluster.

Visualization techniques: Visualization techniques such as box plots and scatter plots can also be used to identify outliers by visually identifying any points that fall outside of the general pattern of the data.

There are many other advanced methods that we will read about in the following section of the article.

Once outliers have been identified, there are several methods for handling them. The appropriate method for handling outliers will depend on the specific dataset and the goals of the analysis. It is important to carefully consider the impact of outliers on the data and the appropriate method for handling them before proceeding with any analysis or modeling. Some common techniques include:

Deleting the outlier observations: This is a simple method, but it can lead to a loss of information if the outliers are actually meaningful observations.

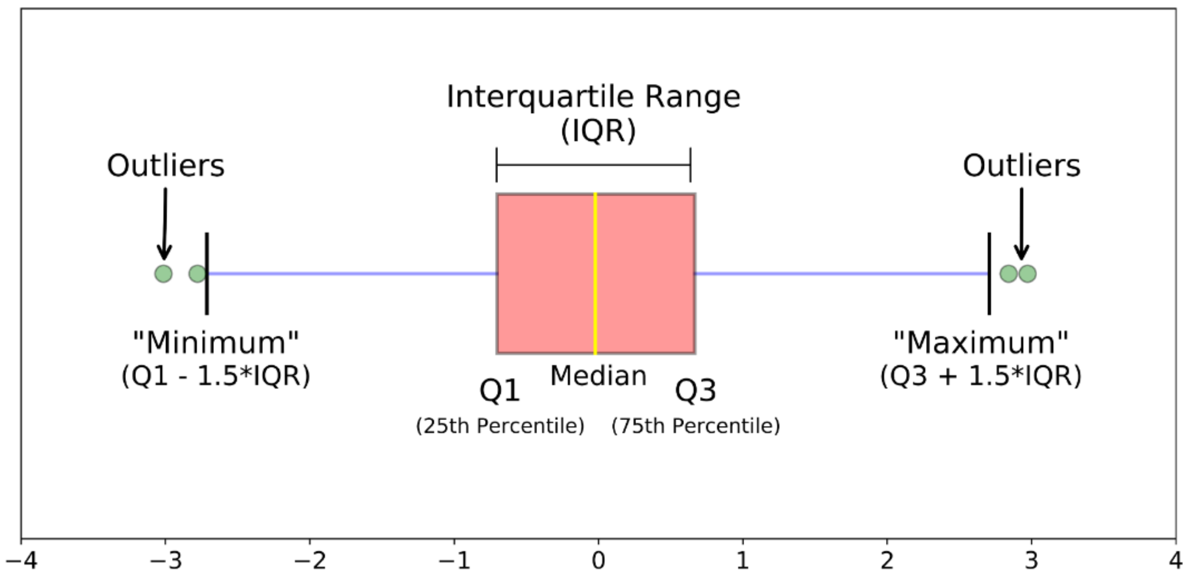
Trimming the data: This method involves removing a certain percentage of the largest and smallest observations.

Winsorizing: This method replaces the outliers with a value that is closer to the center of the data.

Log transformation: This method can be used when the data is positively skewed and the outliers are on the high end of the distribution.

Z-score standardization: This method replaces each observation with its z-score, which is the number of standard deviations away from the mean.

Cap and floor: This method replaces the outlier with a maximum and minimum value respectively.



Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Session 7 & 8

What is Probability Distribution?

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics.

Random experiments are defined as the result of an experiment, whose outcome cannot be predicted. Suppose, if we toss a coin, we cannot predict, what outcome it will appear either it will come as Head or as Tail. The possible result of a random experiment is called an outcome. And the set of outcomes is called a sample point. With the help of these experiments or events, we can always create a probability pattern table in terms of variables and probabilities.

Probability Distribution of Random Variables

A random variable has a probability distribution, which defines the probability of its unknown values. Random variables can be discrete (not constant) or continuous or both. That means it takes any of a designated finite or countable list of values, provided with a probability mass function feature of the random variable's probability distribution or can take any numerical value in an interval or set of intervals. Through a probability density function that is representative of the random variable's probability distribution or it can be a combination of both discrete and continuous.

Two random variables with equal probability distribution can yet vary with respect to their relationships with other random variables or whether they are independent of these. The recognition of a random variable, which means, the outcomes of randomly choosing values as per the variable's probability distribution function, are called **random variates.**

Random Variable

A random variable is a variable whose values can be determined from the outcomes of a random experiment example in an experiment of throwing a pair of dice,

S = {(1,1), (1,2), ……, (6,6), if we define a variable,

X = sum of the numbers on uppermost faces. From the outcomes of sample space we can determine values assume by the variable that is X assumes the values 2,3,….,12 and hence X is a random variable.

According to the values assumed by the random variable we have two tpes of random variable.

- A **discrete random variable** is one whose set of assumed values is countable (arises from counting).
- A **continuous random variable** is one whose set of assumed values is uncountable (arises from measurement.).

Examples:

(1) If the random experiment is throwing an unbiased dice, the random variable associated with this experiment can be defined as the number on the uppermost face, then possible values of X are 1, 2, 3, 4, 5, 6.

(2) If two cards are taken from pack of fifty-two cards, X = number of red cards then X can take the values 0, 1,2.

(3) The number of missed calls received on a particular day.

(4) Suppose milk contents in one liter bags are measured then X = milk content in bag, then X can take values 990 ml $<X \leq 1100$ ml

In examples one and two the sample space of random variable is finite and in third it is countably infinite so these are Discrete random variables where as in fourth sample space interval so values are uncountable hence it is continuous random variable. is

Notation: The random variable is denoted by uppercase letter and its value in lowercase letter. E.g. X denotes a random variable whereas x denotes its value.

Discrete variable: If the number of possible values of the variable is finite or countably infinite then the variable is called as a discrete variable. Thus discrete random variable takes only isolated values.

For example:

(i) Age in completed years.

(ii) Number of arrivals at the Clinic.

(iii) Number of Jobs completed per day.

(iv) Number of accidents on a particular spot per day.

(v) Number of heads in 4 flips of a coin (possible outcomes are 0, 1, 2, 3, 4).

(vi) Number of classes missed last week (possible outcomes are 0, 1, 2, 3, maximum number).

Discrete variables obtained by counting.

PROBABILITY MASS FUNCTION

Let X be a discrete random variable. The set of all possible values of X are denoted by the sample space S. Then probability function

$P(X = x)$ for all S is known as probability mass function if function satisfies the following two conditions.

(i) $0 < P(X = x) < 1$ for all S

(ii) $\Sigma P(X = x) = 1$

CUMULATIVE PROBABILITY DISTRIBUTION FUNCTION

$F(x) = P(X \leq x) = \sum_{-\infty}^{x} P(X = x)$

Ex-1: A random variable X has the following probability distribution values of X.

x	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2 + k$

Find (i) k (ii) $P(X \geq 6)$ (iii) $P(X < 6)$ (iv) $P(0 < X < 5)$

Ex-2: If three coins are tossed simultaneously, find the probability distribution of "number of heads obtained in 3 tosses". Also find the mean, variance and standard deviation of "number of heads obtained in 3 tosses".

Ex-3: In the following table, X is a discrete random variable and p(x) is the probability mass function of X.

x	1	2	3
p(x)	0.3	0.6	0.1

Find the standard deviation and cumulative distribution function of X.

Ex-4: Suppose X represents the minimum of the two numbers when a pair of fair dice is rolled once. Find the probability distribution of X.

Discrete Uniform Distribution

A discrete uniform is the simplest type of all the probability distributions. Discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; such that each value among *n* possible values has equal probability 1/n. In other words discrete uniform distribution could be expressed as "a known, finite number of outcomes equally likely to happen".

Consider the case of throwing a die. The following observations can be made with this action.

(1) The die is numbered from 1 to 6. In probability language we call it as six outcomes.

(2) The die can roll on any number.

(3) The total number of possible events is 6. That is, the die can roll on any of the numbers from 1 to 6. In other words, the events are 'equally likely' with the same probability of 1/6.

(4) The numbers engraved on the die are 6. In other words, the number of values of the random variable is finites.

(5) The numbers are equally spaced or we can say the values of the variables are equally spaced. The experiment leads to discrete Uniform distribution, satisfying all above stated characteristics. Thus a discrete uniform distribution is a probability distribution of equally likely events with equal probability and with finite number of equally spaced outcomes. The discrete uniform distribution is essentially non-parametric, i.e. it does not involve any parameter.

A discrete random variable is said to follow uniform distribution over the range 1,2,2,,,,n if tis pmf is given by

P(X = x) = $\frac{1}{n}$ $\quad x = 1,2,3, \ldots \ldots, n$

= 0 $\quad$ otherwise

Mean = $\frac{n+1}{2}$

Variance = $\frac{n^2-1}{12}$

Binomial Distribution

Bernoulli experiment: Suppose we perform an experiment with two possible outcomes; either success or failure. Success happens with probability p, hence failure occurs with probability q=1-p then such experiment is referred to as Bernoulli experiment.

Probability mass function of Bernoulli variate x is given by

$Px(x) = P(X = x) = p^x\, q^{1-x}\quad x = 0,1;\ 0<p<1; p+q=1$

$\quad = 0$ otherwise $0<p<1 p+q=1$

The binomial experiment means Bernoulli experiment which is repeated n times. The binomial distribution is used to obtain the probability of observing x successes in n trials, with the probability of success on a single trial denoted by p. The binomial distribution assumes that p is fixed for all trials. Here n and p are called as parameters of binomial distribution.

In the above definition notice that the following conditions need to be satisfied for a binomial experiment:

(1) n trials are carried out where n is fixed number.

(2) The outcome of a given trial is either a "success" or "failure".

(3) The probability of success (p) remains constant from trial to trial.

(4) The trials are independent; the outcome of a trial is not affected by the outcome of any other trial.

A discrete random variable X is said to follow **Binomial distribution** with parameters (n,p) if tis pmf given by

$Px(x) = P(X = x) = nCxp^x\, q^{n-x}\quad x = 0,1,2,\ldots.,n;\ 0<p<1; p+q=1$

$\quad = 0$ otherwise $0<p<1 p+q=1$

Mean = np

Variance = npq

Real Life Applications of Binomial Distributions

.

- There are a lot of areas where the application of binomial theorem is inevitable, even in the modern world areas such as computing. In computing areas, binomial theorem has been very useful such as in distribution of IP addresses. With binomial distribution, the automatic allocation of IP addresses is possible.
- Another field that uses Binomial distribution as the important tools is the nation's economic prediction. Economists use binomial theorem to count probabilities to predict the way the economy will behave in the next few years. To be able to come up with realistic predictions, binomial theorem is used in this field.
- Binomial distribution has also been a great use in the architecture industry in design of infrastructure. It allows engineers, to calculate the magnitudes of the projects and thus delivering accurate estimates of not only the costs but also time required to construct them. For contractors, it is a very important tool to help ensuring the costing projects is competent enough to deliver profits.
- The binomial distribution is used when a researcher is interested in the occurrence of an event, not in its magnitude. For instance, in a clinical trial, a patient may survive or die. The researcher studies the number of survivors, and not how long the patient survives after treatment.
- Another example is whether a person is ambitious or not. Here, the binomial distribution describes the number of ambitious persons, and not how ambitious they are.
- Other situations in which binomial distributions arise are quality control, public opinion surveys, medical research, and insurance problems.

Poisson Distribution

The Poisson Distribution was developed by the French mathematician Simeon Denis Poisson in 1837. The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).If we let X = the number of events in a given interval and the mean number of events per interval as λ, then distribution of X is given by Poisson distribution.

A discrete random variable is said to follow Poisson Distribution with parameter λ, if its p.m.f is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2,\ldots.$$

$$= 0 \qquad \text{otherwise}$$

Mean = λ

Variance = λ

Comparison between Binomial and Poisson Distribution

Even though both Binomial and Poisson distribution give probability of X successes, the differences between binomial and Poisson distribution can be drawn clearly on the following grounds:

(1) The binomial distribution is one in which the probability of X successes among n Bernoulli trials is studied. A probability distribution that gives the probability of the count of a number of independent events that occur randomly within a given period, is called Poisson distribution.

(2) Binomial Distribution is biparametric, i.e. it is marked by two parameters n and p whereas Poisson distribution is uniparametric, i.e. described by a single parameter λ.

(3) There are a fixed number of attempts in the binomial distribution. On the other hand, an unlimited number of trials are there in a Poisson distribution.

(4) The success probability is constant in binomial distribution but in Poisson distribution, there are an extremely small number of success.

(5) Poisson distribution can be considered as limiting form of binomial distribution. When success is a rare event i.e. probability of success p is small, $p \to 0$, number of trials is large i.e.

$n \to \infty$, but mean np is finite, binomial distribution tends to Poisson distribution.

(6) In binomial distribution Mean > Variance while in Poisson distribution mean = variance.

Apart from the above differences, there are a number of similar aspects between these two distributions i.e. both are the discrete theoretical probability distribution. Further, on the basis of the values of parameters, both can be unimodal or bimodal. Moreover, the binomial distribution and success probability (p) tends to 0 but A = np is constant. Infinity

Applications of Poisson distribution:

Poisson distribution is applied whenever we observe rare events. Some examples are given below:

- The number of deaths by horse kicking in the Prussian army (first application).
- Birth defects and genetic mutations.
- Rare diseases (like Leukemia, bu3t not AIDS because it is infectious and so not independent). Car accidents.
- Traffic flow and ideal gap distance.
- Number of typing errors on a page. Hairs found in McDonald's hamburgers. Spread of an endangered animal in Africa.
- Failure of a machine in one month.

Geometric Distribution

Geometric distribution is a type of discrete probability distribution that represents the probability of the number of successive failures before a success is obtained in a Bernoulli trial. A Bernoulli trial is an experiment that can have only two possible outcomes, ie., success or failure. In other words, in a geometric distribution, a Bernoulli trial is repeated until success is obtained and then stopped.

The geometric probability distribution is widely used in several real-life scenarios. For example, in financial industries, geometric distribution is used to do a cost-benefit analysis to estimate the financial benefits of making a certain decision. In this article, we will study the meaning of geometric distribution, examples, and certain related important aspects.

What is Geometric Distribution?

The **geometric distribution** is a probability distribution that models the number of trials required to achieve the first success in a sequence of independent Bernoulli trials, where each trial has a constant probability of success.

i.e., Geometric distribution that is based on three important assumptions. These are listed as follows.

- The trials being conducted are independent.
- There can only be two outcomes of each trial - success or failure.
- The success probability, denoted by p, is the same for each trial.

Geometric Distribution Definition

Geometric distribution can be defined as a discrete probability distribution that represents the probability of getting the first success after having a consecutive number of failures. A geometric distribution can have an indefinite number of trials until the first success is obtained.

Geometric Distribution Example

Suppose a dice is repeatedly rolled until "3" is obtained. We know that the probability of getting "3" is p = 1 / 6 and let the random variable X take the values 1, 2, and 3.

- The probability of rolling a 3 in the first trial is 1/6.
- The probability of rolling a 3 in the second trial for the first time is 5/6 × 1/6 = 5/36. Here, 5/6 is the prob of rolling a number that is NOT 3 in the first trial.
- Similarly, the probability of rolling a 3 in the third trial for the first time is, $(5/6)^2$ × 1/6 = 25/216.

Geometric Distribution Formula

There are two geometric probability formulas:

- **Geometric distribution PMF:** $P(X = x) = (1 - p)^{x - 1}p$
- **Geometric distribution CDF:** $P(X \leq x) = 1 - (1 - p)^x$

Formulas for Geometric Distribution

cuemath THE MATH EXPERT

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

$$P(X\leq x) = 1-(1-p)^{x}$$

A geometric distribution can be described by both the probability mass function (PMF) and the cumulative distribution function (CDF). The probability of success of a trial is denoted by p and failure is given by q. Here, q = 1 - p. A discrete random variable, X, that has a geometric probability distribution is represented as X∼G(p)$X\sim G(p)$. Given below are the formulas for the PMF and CDF of a geometric distribution.

Geometric Distribution CDF

The cumulative distribution function of a random variable, X, that is evaluated at a point, x, can be defined as the probability that X will take a value that is lesser than or equal to x. It is also known as the distribution function. The formula for the geometric distribution CDF is given as follows:

$P(X \leq x) = 1 - (1 - p)^x$

Geometric Distribution Mean (Expected Value)

The mean of geometric distribution is also the expected value of the geometric distribution. The expected value of a random variable, X, can be defined as the weighted average of all values of X. The formula for the mean of a geometric distribution is given as follows:

$E[X] = 1 / p$

Variance of Geometric Distribution

Variance can be defined as a measure of dispersion that checks how far the data in a distribution is spread out with respect to the mean. The formula for the variance of a geometric distribution is given as follows:

$Var[X] = (1 - p) / p^2$

Binomial Vs Geometric Distribution

In both geometric distribution and binomial distribution, there can be only two outcomes of a trial, either success or failure. Furthermore, the probability of success will be the same for each trial. The difference between binomial distribution and geometric distribution is given in the table below.

Geometric Distribution	**Binomial Distribution**
A geometric distribution is concerned with the first success only. The random variable, X, counts the number of trials required to obtain that first success.	In a binomial distribution, there are a fixed number of trials and the random variable, X, counts the number of successes in those trials.
The probability mass function is given by PMF = $(1 - p)^{x - 1}p$	The probability mass function is given by PMF = $({}_{nx})p_{x}(1-p)_{n-x}(nx)px(1-p)n-x$
Mean = 1 / p, Variance = $(1 - p) / p^2$	Mean = np, Variance = np(1-p)

Important Notes on Geometric Distribution

- The geometric distribution is a discrete probability distribution where the random variable indicates the number of Bernoulli trials required to get the first success.
- The probability mass function of a geometric probability distribution is $(1 - p)^{x - 1}p$ and the cumulative distribution function is $1 - (1 - p)^x$.
- The mean of a geometric distribution is 1 / p and the variance is $(1 - p) / p^2$.

Continuous Random Variable

Continuous variable: A variable which assumes infinitely many values included in some range is defined as continuous random variable.

We call X a continuous random variable in a ≤ x ≤ b, if X can take on any value in this interval. An example of a random variable is the height of a person; say an adult male, selected randomly from a population. (This height typically takes on values in the range 1.64 < x < 9.84 feet, say, so a = 1.64 and b = 9.84.) If we select a male at random from a large population, and measure his height, the measured height can take on any real number within the interval of interest. This compels us to redefine our idea of a distribution, related to continuous variable, using a continuous function in place of the discrete function to represent probability of continuous variable.

A Continuous variable takes all possible values in a range set which is in the form of interval. On the other hand discrete random variable takes only specific or isolated values. Generally values of discrete random variable are obtained by counting while values of continuous variable are obtained by measuring to any degree of accuracy. The values of continuous variable move continuously from one possible value to another, without having to jump as in case of a discrete variable. E.g. Random variable X which is number of F.Y.B.Sc. students is discrete. While random variable Y which is height of a student is continuous variable. Continuous Variables would (literally) take forever to count. In fact, you would get to "forever" and never finish counting them. For example, take age. You can't count "age". Why not? Because it would literally take forever. For example, you could be: 20 years, 10 months, 2 days, 5 hours, 10 minutes, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picoseconds...and so on. Even though age is a continuous could turn age into a discrete variable and then you could count it. For example:

- A person's age in years.
- A baby's age in months.

Hence age, height, weight, time, income are continuous variables but we can turn them into discrete variable by appropriate definition.

Examples of continuous random variable:

- Consumption of cooking oil in a house.
- Waiting time on a ticket window.
- Life of an electronic gadget.
- Yield of a crop in certain area.
- Temperature in Mumbai
- Time it takes a computer to complete a task.

Properties of pdf of continuous random variable

i) f(x)>0, for all x belongs to sample Space S since probability >0

ii) $\int_{-\infty}^{\infty} f(x)df = 1$ implied that total area bounded by the curve of density function and X – axis equal to 1, when computed over entire range of variable X.

Continuous Uniform Distribution

The simplest continuous probability distribution is the Uniform distribution also known as Continuous Uniform distribution. It is observed in many situations. In general, if the probabilities of various classes of a continuous variable are more or less the same, the situation is best described by uniform distribution. In this distribution the p.d.f. of random variable remains constant over the range space of variable.

Definition:

A continuous random variable is said to follow uniform distribution over the interval (a, b), if its p.d.f. is given by

$$f(x) = \frac{1}{b-a}; \quad a \leq x \leq b$$

$$= 0 \qquad \text{otherwise}$$

And is represented as X~ U(a, b)

Mean = $\frac{a+b}{2}$

Variance = $\frac{(b-a)^2}{12}$

Normal Distribution

Normal distribution is an important continuous distribution because a good number of random variables occurring in practice can be approximated to it. If a random variable is affected by many independent causes, and the effect of each cause is not overwhelmingly large as compared to other effects, then the random variable will closely follow a normal distribution.

Pioneers of Normal distribution:

Normal distribution was first mentioned by De-Moivre in 1733 and was also known to Laplace in 1774. Independently, the mathematicians Adrain in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors in astronomy were fit well by this distribution. Quételet and Galton were the first to apply the normal distribution to human as well as animal characteristics. He noted that characteristics such as height, weight, and strength were normally distributed. Anyhow, the credit of Normal distribution has been given to Gauss and is often called as Gaussian distribution. Normal distribution is the maximally used distribution in the theory of Statistics.

Definition:

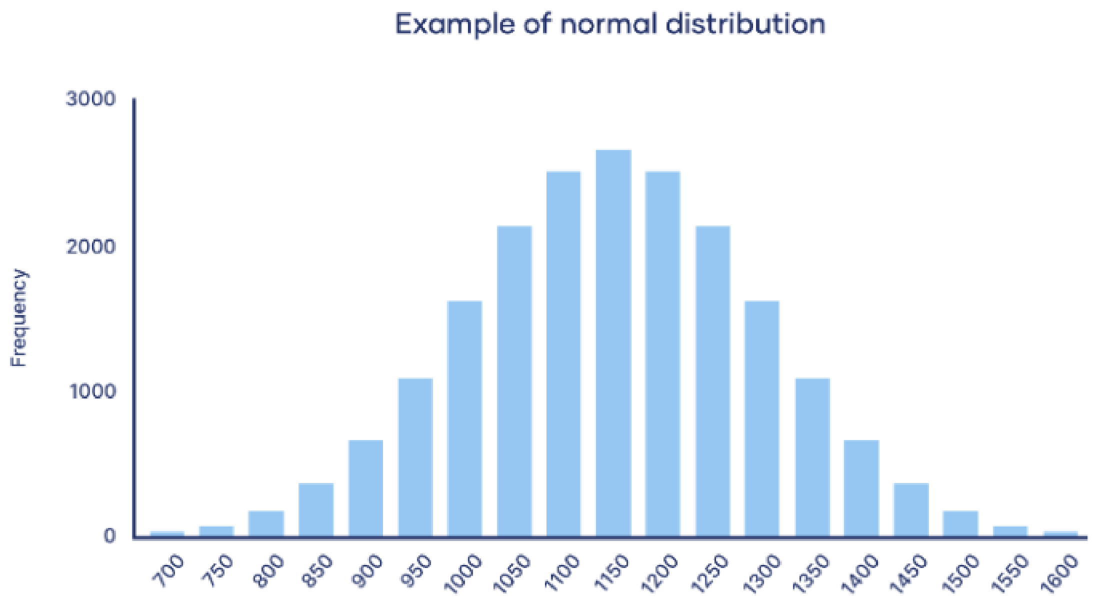
A continuous random variable X is said to follow normal distribution with parameters μ and o^2, if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; \quad -\infty < x < \infty; -\infty < \mu < \infty; \sigma > 0$$

It is denoted as X~N (μ, σ^2)

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

Normal distributions are also called Gaussian distributions or bell curves because of their shape.



Why do normal distributions matter?

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

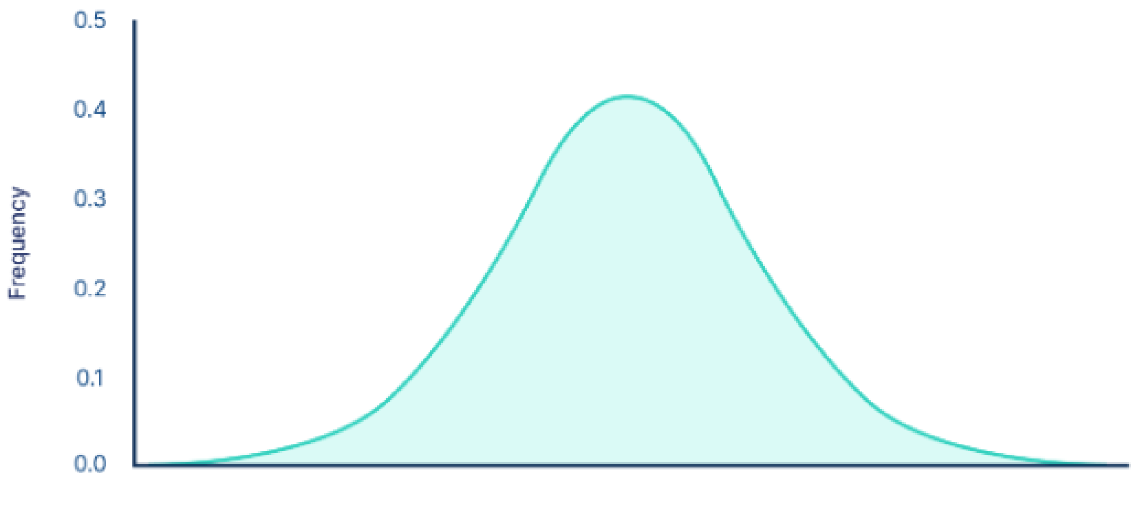
Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.

Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

What are the properties of normal distributions?

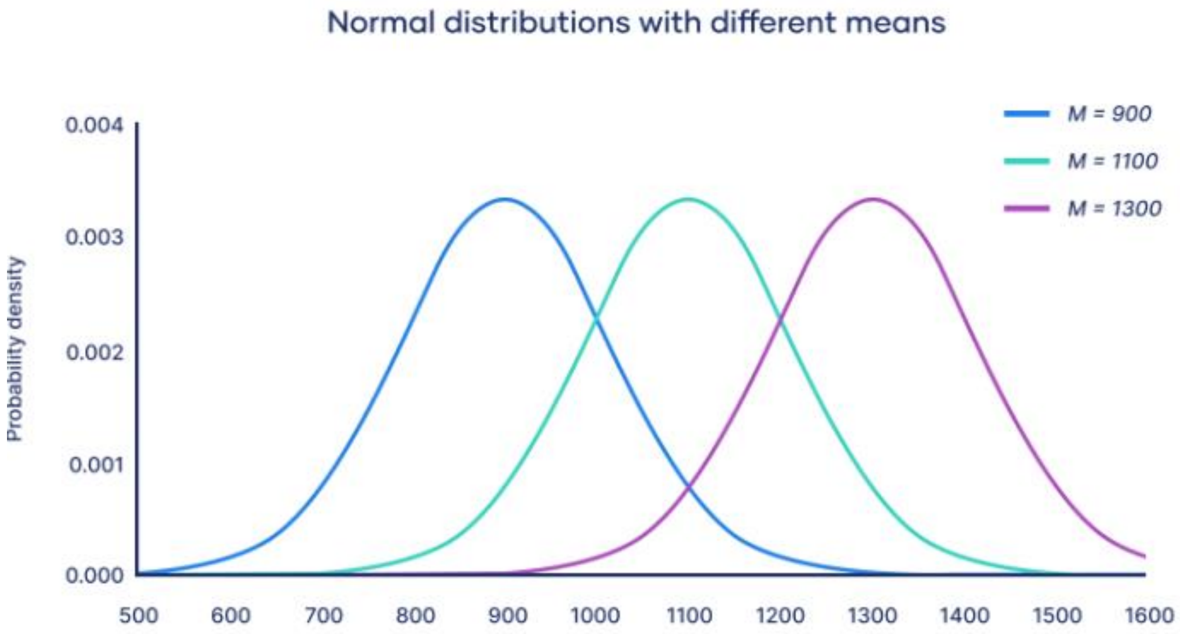
Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

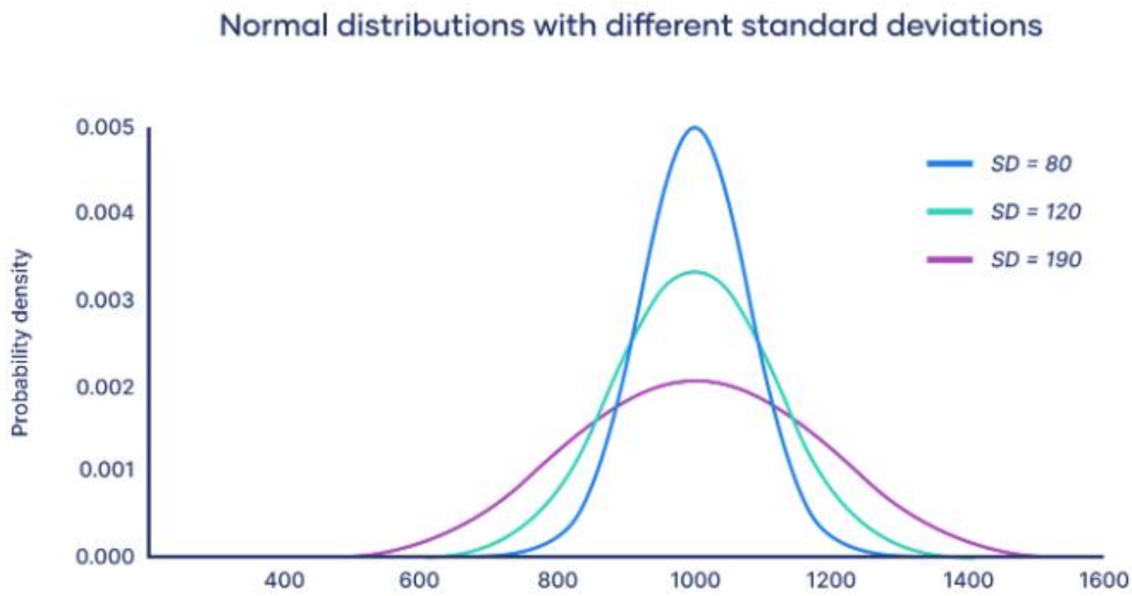


The mean is the location parameter while the standard deviation is the scale parameter.

The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.



The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



Empirical rule

The **empirical rule**, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.

Central limit theorem

The central limit theorem is the basis for how normal distributions work in statistics.

In research, to get a good idea of a population mean, ideally you'd collect data from multiple random samples within the population. A **sampling distribution of the mean** is the distribution of the means of these different samples.

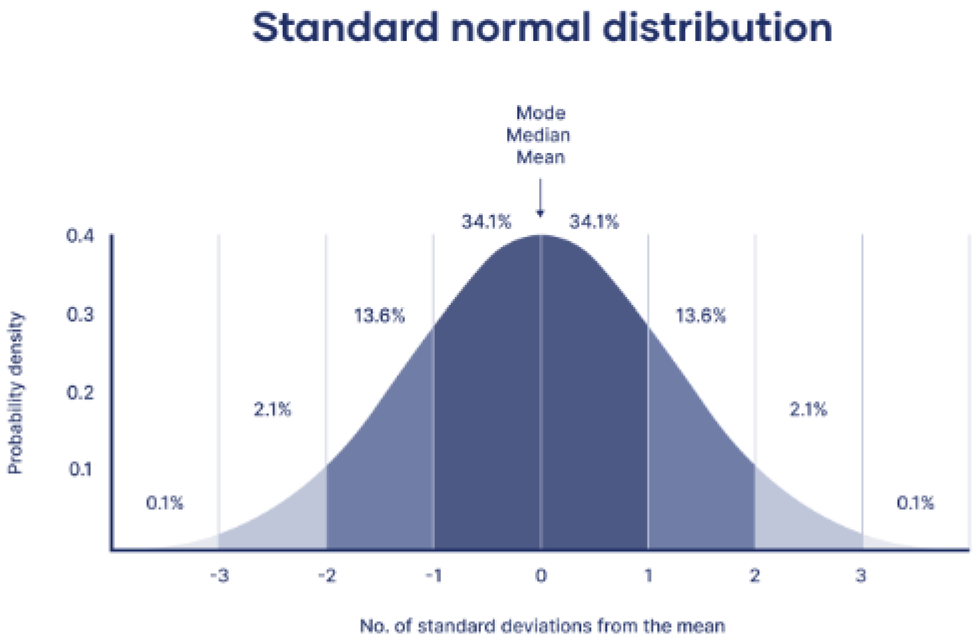
The central limit theorem shows the following:

- Law of Large Numbers: As you increase sample size (or the number of samples), then the sample mean will approach the population mean.
- With multiple large samples, the sampling distribution of the mean is normally distributed, even if your original variable is not normally distributed.

Parametric statistical tests typically assume that samples come from normally distributed populations, but the central limit theorem means that this assumption isn't necessary to meet when you have a large enough sample.

You can use parametric tests for large samples from populations with any kind of distribution as long as other important assumptions are met. A sample size of 30 or more is generally considered large.

For small samples, the assumption of normality is important because the sampling distribution of the mean isn't known. For accurate results, you have to be sure that the population is normally distributed before you can use parametric tests with small samples.



You only need to know the mean and standard deviation of your distribution to find the *z*-score of a value.

Z-score Formula	Explanation
$z = \frac{x - \mu}{\sigma}$	• *x* = individual value • μ = mean • σ = standard deviation

EXPONENTIAL DISTRIBUTION

In Probability theory and statistics, the exponential distribution is a continuous probability distribution that often concerns the amount of time until some specific event happens. It is a process in which events happen continuously and independently at a constant average rate. The exponential distribution has the key property of being memoryless. The exponential random variable can be either more small values or fewer larger variables. For example, the amount of money spent by the customer on one trip to the supermarket follows an exponential distribution.

Exponential Distribution Formula

The continuous random variable, say X is said to have an exponential distribution, if it has the following probability density function:

The continuous random variable, say X is said to have an exponential distribution, if it has the following probability density function:

$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & for\ x \leq 0 \end{cases}$$

Where

λ is called the distribution rate.

Memorylessness: The exponential distribution exhibits the memoryless property, meaning that the probability of an event occurring in the next instant does not depend on how much time has already elapsed.

Mean and Variance:

The mean and variance of the exponential distribution are given by:

- **Mean** (μ): $E[X] = \frac{1}{\lambda}$
- **Variance** (σ^2): $Var[X] = \frac{1}{\lambda^2}$

Session 9 & 10

Descriptive Statistical measures

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5). The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set.

Central Tendency

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analysed data.

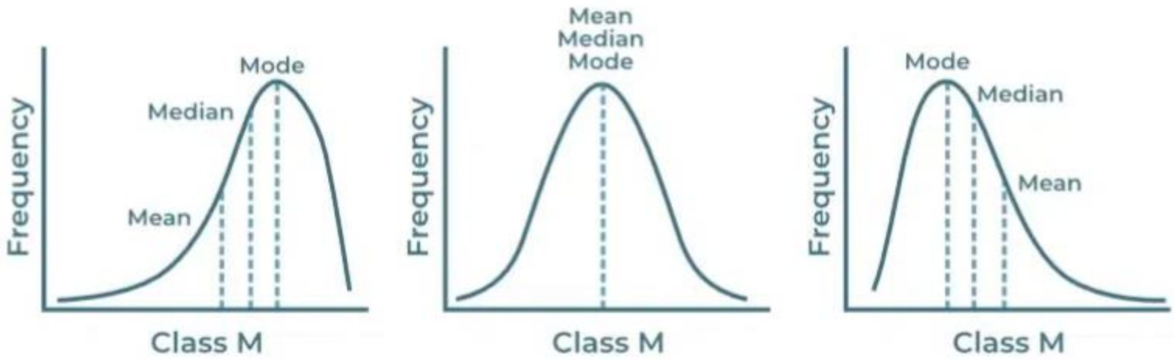
Measures of central tendency describe the centre position of a distribution for a data set. A person analyses the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analysed data set.

Measures of Variability

Measures of variability (or the measures of spread) aid in analysing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).



Measures of Central Tendency (Averages)

One of the most important objectives of statistical analysis is to get one single value that describes the characteristic of the entire data. An average is the value of the variable which is representative of the entire data. It gives us idea about the concentration of the values in the central part of the distribution.

Types of Averages:

1) Arithmetic Mean (A.M.)
2) Weighted Arithmetic Mean
3) Median
4) Mode
5) Geometric Mean (G.M.)
6) Harmonic Mean (H.M.)

Requisites of a good average:

1) Easy to understand
2) Simple to compute
3) Based on all the items.
4) Not unduly affected by extreme observations.
5) Rigidly defined.
6) Capable for further algebraic treatment.
7) Easy to interpret

Arithmetic Mean(A.M.):

Arithmetic mean for Raw Data:

The A.M. of N observations $X_1, X_2, \ldots, X_N$ is denoted by $\bar{X}$ and is defined as follows.

$$\text{A.M.} = \overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X}{N}$$

Ex: The grades of a student on six examinations were 84, 91, 72, 68, 87 and 78. Find the arithmetic mean of the grades.

Ex: The mean of 10 observations was found to be 20. Later on it was discovered that the observations 24 and 34 were wrongly noted as 42 and 54. Find the corrected mean.

Arithmetic Mean for a frequency distribution:

Consider a data of n observations $X_1, X_2, \ldots, X_n$ occurring with respective frequencies $f_1, f_2, \ldots, f_n$. Then the A.M. is denoted by $\bar{X}$ and is defined as follows.

$$\text{A.M.} = \overline{X} = \frac{f_1X_1 + f_2X_2 + \ldots + f_nX_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum fX}{\sum f}$$

We denote Σf by N, called as total frequency

$$\therefore \text{A.M.} = \overline{X} = \frac{\sum fX}{N}$$

Ex: The following table gives the monthly income of 20 families in a city. Calculate the arithmetic mean.

Income (in '00 Rs.)	16	20	30	35	45	50
No. of families	2	5	4	6	2	1

Ex: Use the following frequency distribution of heights to find the arithmetic mean of height of 100 students at XYZ university.

Height (inches)	Number of students
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8

Ex: Use the following frequency distribution of weekly wages to find the arithmetic mean of wage of employees at P & R company.

Weekly Wage ($)	Number of employees
250.00 – 259.99	8
260.00 – 269.99	10
270.00 – 279.99	16
280.00 – 289.99	14
290.00 – 299.99	10
300.00 – 309.99	5
310.00 – 319.99	2

Ex: Find the missing frequency if the mean is 21.9

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	5	--	13	21	16	8	3

Combined Mean:

Ex: The average marks of a group of 100 students in Statistics are 60 and for other group of 50 students, the average marks are 90. Find the average marks of the combined group of 150 students.

Answer: Given: $N_1 = 100$, $\overline{X}_1 = 60$, $N_2 = 50$, $\overline{X}_2 = 90$

Combined Mean = $\overline{X} = \dfrac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$

$$= \frac{(100 \times 60) + (50 \times 90)}{100 + 50} = \underline{\mathbf{70}}$$

Weighted Arithmetic Mean:

Sometimes we associate certain weighing factor (or weights) with the numbers X_1, $X_2, \ldots, X_n$. Suppose the weights are $w_1, w_2, \ldots, w_n$. The weighted arithmetic mean is denoted by $\bar{X}$ and is defined as follows.

$$\text{Weighted A.M.} = \vec{X} = \frac{w_1X_1 + w_2X_2 + \ldots + w_nX_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\sum wX}{\sum w}$$

Ex: A student's grades in laboratory, lecture and recitation parts of a Physics course were 71, 78 and 89 respectively. If the weights of these grades are 2, 4 and 5 respectively, what is the appropriate average grade?

Answer:

X	w	wX
71	2	142
78	4	312
89	5	445
	Σw = 11	ΣwX = 899

$$\text{Weighted A.M.} = \vec{X} = \frac{\sum wX}{\sum w} = \frac{899}{11} = 81.7273$$

MERITS AND DEMERITS OF MEAN

Merits:

1. It is rigidly defined.

2.It is easy to understand and easy to calculate.

3.It is based on all the observations.

4.It is capable of further algebraic treatment.

5.Of all the averages, A.M. is least affected by sampling fluctuations i.e. it is a stable average.

Demerits:

1. It cannot be obtained by mere inspection nor can it be located graphically.

2. It cannot be obtained even if a single observation is missing. It is affected by extreme values.

3. It is affected by extreme values

4.It cannot be calculated for frequency distribution having open end class- intervals e.g. class-intervals like below 10 or above 50 etc.

5.It may be a value which may not be present in the data.

6. Sometimes, it gives absurd results. e.g. Average number of children per family is 1.28.

7.It cannot be used for the study of qualitative data such as intelligence, honesty, beauty etc.

Even though A.M. has various demerits, it is considered to be the best all averages as it satisfies most of the requisites of a good average. A.M. is called the Ideal Average.

Median:

The **Median** of a set of numbers arranged in order of magnitude is either the middle value or arithmetic mean of the two middle values.

Median for raw data:

Suppose there are N observations.

If N is an odd number,

median = $\left(\frac{N+1}{2}\right)^{th}$ observation in order sample

If N is an even number,

median = average of $\left(\frac{N}{2}\right)^{th}$ & $\left(\frac{N}{2}+1\right)^{th}$ observation in order sample

Ex: The number of ATM transactions per day was recorded at 15 locations in a large city. The data were as follows.

35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32, 60.

Find the median number of transactions.

Answer: Given observations in ascending order:

30, 32, 35, 40, 47, 48, 49, **50**, 52, 55, 60, 65, 76, 225, 325

N = 15, which is an 'odd' number

$\left(\frac{N+1}{2}\right) = 8$

median = $\left(\frac{N+1}{2}\right)^{th}$ observation in order sample

= 8^{th} observation in order sample = **50**

Ex: The following table gives the data of weight of 20 students at a University. Find the median weight.

138, 146, 168, 146, 161, 164, 158, 126, 173, 145,

150, 140, 138, 142, 135, 132, 147, 176, 147, 142

Ex: Find the median of the following data.

X	10	15	25	40	60	75
f	3	4	6	17	12	7

Answer: $\Sigma f = N = 49$, N is an odd number

x	10	15	25	40	60	75
f	3	4	6	17	12	7
less than cumulative freq	3	7	13	30	42	49

median = $\left(\dfrac{N+1}{2}\right)^{th}$ observation in order sample

= 25^{th} observation in order sample = **40**

Median for grouped data:

Suppose the total number of observations is N. The "median class" is defined as the class interval for which the less than cumulative frequency is just greater than N/2. Then,

median = $\ell_1 + \left\{\dfrac{(\ell_2 - \ell_1)}{f}\left(\dfrac{N}{2} - \text{c.f.}\right)\right\}$

ℓ_1 = lower class boundary of median class

ℓ_2 = upper class boundary of median class

f = frequency of median class

c.f. = less than cumulative frequency of class preceding median class

<u>Ex:</u> Find the median of the following data.

Weight (pounds)	Frequency
118 – 126	3
127 – 135	5
136 – 144	9
145 – 153	12
154 – 162	5
163 – 171	4
172 – 180	2

<u>Answer:</u> Given: N = Σf = 40 ; N/2 = 20

Weight (pounds)	Frequency	Less Than Cumulative Frequency
118 – 126	3	3
127 – 135	5	8
136 – 144	9	**<u>17</u>**
<u>145 – 153</u>	**<u>12</u>**	29
154 – 162	5	34
163 – 171	4	38
172 – 180	2	40

The median class is 145 – 153

$$\text{median} = \ell_1 + \left\{\frac{(\ell_2 - \ell_1)}{f}\left(\frac{N}{2} - \text{c.f.}\right)\right\}$$

$$= 144.5 + \left[\left(\frac{153.5 - 144.5}{12}\right) \times (20 - 17)\right] = \underline{\mathbf{146.75}}$$

MERITS AND DEMERITS OF MEDIAN

Merits:

1.It is easy to understand and easy to calculate.

2.It is quite rigidly defined.

3.It can be computed for a distribution with open-end classes.

4.In majority of the cases, it is one of the values in the data.

5.It can be determined graphically.

6.Since median is a positional average, it can be computed even if the observations at the extremes are unknown.

7.It is not highly affected by fluctuations in sampling.

8.It can be calculated even for qualitative data.

Demerits:

1.When the number of observations is large, the pre-requisite of arranging observations in ascending/descending order of magnitude is a difficult process.

2. It is not based on all observations and hence, may not be a proper representative.

3. It is not capable of further mathematical treatment.

4. Since it does not require information about all the observation, it is insensitive to some changes

Mode:

The **Mode** of a set of numbers is that value which occurs with the greatest frequency, that is, it is the most common value.

Note:

1) Sometimes the **Mode** may not exist.
2) Even if the **Mode** exists, sometimes it may not be unique.

Ex: The reaction times of an individual to a certain stimulus were measured as 0.53, 0.46, 0.50, 0.49, 0.52, 0.44, 0.55, 0.53, 0.40 and 0.56. Find the mode.

Ex: Three teachers of a subject reported examination grade of 79, 74 and 82 in their classes, which consisted of 32, 25 and 17 students respectively. Determine the mode.

Mode for a grouped data:

"Modal class" is the class interval with maximum frequency.

$$\text{mode} = \ell_1 + \left\{ (\ell_2 - \ell_1) \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \right\}$$

ℓ_1 = lower class boundary of modal class

ℓ_2 = upper class boundary of modal class

f_0 = frequency of the class preceding modal class
f_1 = frequency of modal class
f_2 = frequency of the class next to modal class

MERITS AND DEMERITS OF MODE

Merits:

1. It is easy to understand and simple to calculate.

2.It is not affected by extreme values or sampling fluctuations.

3.It can be calculated for distribution with open-end classes.

4.It can be determined graphically.

5.It is always present within the data and is the most typical value of the given set of data.

6. It is applicable to both, qualitative and quantitative data.

Demerits

1.It is not rigidly defined.

2.It is not based on all observations.

3.It is not capable of further mathematical treatment.

4.It is indeterminate if the modal class is at the extreme of the distribution.

5. If the sample of data for which mode is obtained is small, then such mode has no significance.

Geometric Mean (G.M.):

Geometric Mean for Raw Data:

The G.M. of N observations $X_1, X_2, \ldots, X_N$ is denoted by G and is defined as follows.

G.M. = G = $\sqrt[N]{X_1\, X_2 \ldots X_N}$

Ex: Find the Geometric Mean of the following data.

28.5, 73.6, 47.2, 31.5 and 64.8.

Answer: N = 5

G.M. = G = $\sqrt[N]{X_1\, X_2 \ldots X_N}$

= $\sqrt[5]{28.5 \times 73.6 \times 47.2 \times 31.5 \times 64.8}$ = **45.8258**

Geometric Mean for a frequency distribution:

Consider a data of n observations $X_1, X_2, \ldots, X_n$ occurring with respective frequencies $f_1, f_2, \ldots, f_n$. Then the G.M. is denoted by G and is defined as follows.

G.M. = G = $\sqrt[N]{(X_1)^{f_1}\,(X_2)^{f_2} \ldots (X_n)^{f_n}}$ where $N = \Sigma f$

Harmonic Mean(H.M.):

Harmonic mean for Raw Data:

The H.M. of N observations $X_1, X_2, \ldots, X_N$ is denoted by H and is defined as follows.

$$H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}}$$

$$\therefore \frac{1}{H} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}}{N}$$

$$\therefore \frac{1}{H} = \frac{1}{N}\left(\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}\right)$$

$$\therefore \frac{1}{H} = \frac{1}{N}\sum\left(\frac{1}{X}\right)$$

Ex: Cities A, B and C are equidistant from each other. A motorist travels from A to B at 30 mph, from B to C at 40 mph and from C to A at 50 mph. Find his average speed.

Answer:

$$\frac{1}{H} = \frac{1}{N}\sum\left(\frac{1}{X}\right) = \frac{1}{3}\left(\frac{1}{30} + \frac{1}{40} + \frac{1}{50}\right) = \frac{47}{1800}$$

$\therefore$ H = Harmonic Mean = $\frac{1800}{47}$ = **38.2979**

Harmonic Mean for a frequency distribution:

Consider a data of n observations $X_1, X_2, \ldots, X_n$ occurring with respective frequencies $f_1, f_2, \ldots, f_n$. Then the H.M. is denoted by H and is defined as follows.

$$H = \frac{N}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \ldots + \frac{f_n}{X_n}} \quad \text{where } N = \Sigma f$$

Ex: An airplane travels distances of 2500, 1200 and 500 miles at speeds 500, 400 and 250 mph respectively. Find the Harmonic mean.

Answer:

$$\text{Harmonic Mean} = \frac{\sum w}{\sum\left(\frac{w}{X}\right)}$$

$$= \frac{2500 + 1200 + 500}{\left(\frac{2500}{500}\right) + \left(\frac{1200}{400}\right) + \left(\frac{500}{250}\right)} = \underline{\mathbf{420}}$$

Quartiles, Deciles and Percentiles:

Suppose we arrange a set of data in order of magnitude. The values which divide the set into four equal parts are denoted by Q_1, Q_2, Q_3 and are called as the 1st, 2nd, 3rd Quartiles respectively. Similarly, the values which divide the set into ten equal parts are denoted by $D_1, D_2, \ldots, D_9$ and are called as the 1st, 2nd, ..., 9th Deciles respectively. And, the values which divide the set into hundred equal parts are denoted by $P_1, P_2, \ldots, P_{99}$ and are called as the 1st, 2nd, ..., 99th Percentiles respectively. Collectively, Quartiles, Deciles and Percentiles are called as **Quantiles**.

Q_i = i^{th} Quartile = $L_1 + \left\{\frac{c}{f}\left(\frac{iN}{4} - c.f.\right)\right\}$ where $i = 1, 2, 3$

L_1 = lower class boundary of i^{th} quartile class
f = frequency of i^{th} quartile class
c.f. = less than cumulative freq of class preceding i^{th} quartile class

c = class width

D_i = i^{th} Decile = $L_1 + \left\{\frac{c}{f}\left(\frac{iN}{10} - c.f.\right)\right\}$ where $i = 1, 2, \ldots, 9$

L_1 = lower class boundary of i^{th} decile class
f = frequency of i^{th} decile class
c.f. = less than cumulative freq of class preceding i^{th} decile class

c = class width

P_i = i^{th} Percentile = $L_1 + \left\{\frac{c}{f}\left(\frac{iN}{100} - c.f.\right)\right\}$ where $i = 1, 2, \ldots, 99$

L_1 = lower class boundary of i^{th} percentile class
f = frequency of i^{th} percentile class
c.f. = less than cumulative freq of class preceding i^{th} percentile class

c = class width

Ex: For the following data of age, calculate the 1st Quartile, 5th Decile and 54th Percentile.

Age in years	Number of persons
0 – 10	6
10 – 20	8
20 – 30	13
30 – 40	18
40 – 50	16
50 – 60	13
60 – 70	12
70 – 80	9
80 – 90	4
90 – 100	1

Answer: $N = \Sigma f = 100$

Class Interval	Frequency	Less Than Cumulative Frequency
0 – 10	6	6
10 – 20	8	14
20 – 30	13	27
30 – 40	18	45
40 – 50	16	61
50 – 60	13	74
60 – 70	12	86
70 – 80	9	95
80 – 90	4	99
90 – 100	1	100

1) **To find Q_1:**

Consider $\frac{iN}{4}$, put i = 1

$\frac{1 \times N}{4}$ = 25, class containing Q_1 is 20-30

Q_i = i^{th} Quartile = $L_1 + \left\{\frac{c}{f}\left(\frac{iN}{4} - c.f.\right)\right\}$, put i = 1

$Q_1 = L_1 + \left\{\frac{c}{f}\left(\frac{1 \times N}{4} - c.f.\right)\right\} = 20 + \left\{\frac{10}{13}(25 - 14.)\right\}$ = **28.4615**

2) **To find D_5:**

Consider $\frac{iN}{10}$, put i = 5

$\frac{5 \times N}{10}$ = 50, class containing Q_1 is 40-50

D_i = i^{th} Decile = $L_1 + \left\{\frac{c}{f}\left(\frac{iN}{10} - c.f.\right)\right\}$, put i = 5

$D_5 = L_1 + \left\{\frac{c}{f}\left(\frac{5N}{10} - c.f.\right)\right\} = 40 + \left\{\frac{10}{16}(50 - 45.)\right\}$ = **43.125**

3) **To find P_{54}:**

Consider $\frac{iN}{100}$, put i = 54

$\frac{54 \times N}{100}$ = 54, class containing Q_1 is 40-50

$P_i = i^{th}$ Percentile = $L_1 + \left\{ \frac{c}{f}\left(\frac{iN}{100} - c.f. \right) \right\}$, put i = 54

$P_{54} = L_1 + \left\{ \frac{c}{f}\left(\frac{54N}{100} - c.f. \right) \right\}$ = 40 +

$\left\{ \frac{10}{16}(54 - 45.) \right\}$ = **45.625**

Comparison between Central Tendencies

We have studied five different measures of central tendency. It is obvious that no single measure can be the best for all situations. The most commonly used measures are mean, median and mode. It is not desirable to consider any one of them to be superior or inferior in all situations. The selection of appropriate measure of central tendency would largely depend upon the nature of the data; more specifically, on the scale of measurement used for representing the data and the purpose on hand.

The data obtained on nominal scale, we can count the number of cases in each category and obtain the frequencies. We may then be interested in knowing the class which is most popular or the most typical value in the data. In such cases, mode can be used as the appropriate measure of central tendency.

e.g. Suppose in a genetical study, for a group of 50 family members, we want to know most common colour of eyes. Then we count the number of persons for each different colour of eye. Suppose 3 persons have light eyes, 6 persons have brown eyes, 12 with dark grey eyes and 29 persons are with black eyes. Then the most common colour of eyes (i.e. mode) for this group of people is 'black'.

When the data is available on ordinal scale of measurement i.e. the data is provided in rank order, use of median as a measure of central tendency is appropriate. Suppose in a group of 75 students, 10 students have failed, 15 get pass class, 20 secure second class and 30 are in first class. The average performance of the students will be the performance of the middlemost student (arranged as per rank) i.e. the performance of 38th student i.e. second class; which is the median of the data. Median is only a point on the scale of measurement, below and above which lie exactly 50% of the data. Median can also be used (i) for truncated (incomplete) data, provided we know the total number of cases and their positions on the scale and (ii) when the distribution is markedly skewed.

Arithmetic Mean is the most commonly used measure of central tendency. It can be calculated when the data is complete and is represented on interval or ratio scale. It represents the centre of gravity of the data i.e. the measurements in any sample are perfectly balanced about the mean. In computation of simple A.M., equal importance is given to all observations in the data. It is preferred because of its high reliability and its applicability to inferential statistics. Thus, A.M. is more precise, reliable and stable measure of central tendency.

What are the objectives of computing dispersion?

(1) Comparative study

- Measures of dispersion give a single value indicating the degree of consistency or uniformity of distribution. This single value helps us in making comparisons of various distributions.
- The smaller the magnitude (value) of dispersion, higher is the consistency or uniformity and vice-versa.

(2) Reliability of an average

- A small value of dispersion means low variation between observations and average. It means that the average is a good representative of observation and very reliable.
- A higher value of dispersion means greater deviation among the observations. In this case, the average is not a good representative and it cannot be considered reliable.

(3) Control the variability

- Different measures of dispersion provide us data of variability from different angles, and this knowledge can prove helpful in controlling the variation.
- Especially in the financial analysis of business and medicine, these measures of dispersion can prove very useful.

(4) Basis for further statistical analysis

- Measures of dispersion provide the basis for further statistical analysis like computing correlation, regression, test of hypothesis, etc.

Range

Definition: If L is the largest observation in the data and S is the smallest observation, then range is the difference between L and S. Thus,

Range = L-S

For a frequency distribution, range may be considered as the difference between the largest and the smallest class-boundaries.

Range is a crude and simplest measure of dispersion. It measures the scatter of observations among themselves and not about any average.

The corresponding relative measure is

Coefficient of range = $\frac{L-S}{L+S}$

Note:

1.Range is a suitable measure of dispersion in case of small groups. In the branch of statistics known as Statistical Quality Control, range is widely used. It is also used to measure the changes in the prices of shares. Variation in daily temperatures at a certain place are measured by recording maximum temperature and minimum temperature. Range is also used in medical sciences to check whether blood pressure, haemoglobin count etc. are normal.

2. The main drawback of this measure is that it is based on only two extreme values, the maximum and the minimum, and completely ignores all the remaining observations.

Quartile Deviation

We have seen earlier that range, as a measure of dispersion, is based only on two extreme values and fails to take into account the scatter of remaining observations within the range. To overcome this drawback to an extent, we use another measure of dispersion called Inter-Quartile Range. It represents the range which includes middle 50% of the distribution. Hence,

Inter-Quartile Range = $Q_3 - Q_1$

where, Q_3 and Q_1 represent upper and lower quartiles respectively.

Half of Inter-Quartile-Range i.e. Semi-Inter-Quartile Range = $\frac{Q_3 - Q_1}{2}$ is also used as absolute measure of dispersion. The semi-inter-quartile range is popularly known as Quartile Deviation (Q.D.)

$$\boldsymbol{QD = \frac{Q_3 - Q_1}{2}}$$

The corresponding relative measure of dispersion is called coefficient of quartile deviation and is defined as

Coefficient of Q.D = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Note:

1. Q.D. is independent of extreme values. It is a better representative and more reliable than range.

2. Q.D. gives an idea about the distribution of middle half of the observations around the median.

3. Whenever median is preferred as a measure of central tendency, quartile deviation is preferred as a measure of dispersion. However, like median, quartile deviation is also not capable of further algebraic treatment, as it does not take into consideration all the values of the distribution.

4. For a symmetric distribution,

Q_1 = Median - Q.D. and Q_3 = Median + Q.D.

STANDARD DEVIATION (S.D.)

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical techniques.

Definition:

Standard Deviation (S.D.) is defined as the positive square root of the arithmetic mean of the squares of the deviations of the observations from their arithmetic mean.

The arithmetic mean of the squares of the deviations of the observations from their A.M. is called variance

Thus SD = $+\sqrt{variance}$

Note: The coefficient of variation is considered to be the most appropriate measure for comparing variability of two or more distributions. The importance of C.V. can be explained with the following example: Suppose milk bags are filled with automatic machine, the amount of milk being 1 litre per bag. Setting the machine for C.V. = 0 (i.e. zero variability) is impossible, because of chance causes which exist in any process and are beyond human control. Let us assume that the machine is set for C.V. less than or equal to 1. Then, using statistical law, we can expect approximately 99.73% of the bags to contain milk quantity ranging between atleast 970 mls. and at the most 1030 mls. Usually, this variation is not noticable and hence acceptable to customers But, if the machine is set for C.V. say equal to 5, we can see that about 16% of the bags will contain 900 mls. or less of milk, which is definitely not acceptable. Thus, one has to take utmost care to reduce C.V.

In manufacturing process, with reference to quality control section and in pharmaceutical industries, C.V. plays a very important role. In quality control section, efforts are made to improve the quality by producing items as per given specifications. The extent of deviation from given specifications can be measured using C. V. The lower is the value of C.V., better is the quality of the items produced. Due to competition, almost all industries have reduced the C.V. of their goods to a considerable extent in last few years.

In pharmaceutical industries, C.V. is as low as 1 or less than 1. The variation in the weights of tablets is almost negligible.

In industrial production, C.V. depends upon raw material used. A good quality of raw material will result in homogeneous end product. In chemical and pharmaceutical industries, C.V. can be reduced by thorough mixing and pounding of the raw material.

MERITS AND DEMERITS OF STANDARD DEVIATION

Merits:

1.It is rigidly defined.

2.It is based on all observations.

3.It is capable of further algebraic treatment.

4.It is least affected by sampling fluctuations.

Demerits:

1.As compared to other measures it is difficult to calculate.

2.It cannot be calculated for distribution with open-end class-intervals.

3.It gives more importance (weightage) to extreme values and less importance to the values close to A.M. that is unduly affected due to extreme observations

4.It cannot be calculated for qualitative data.

Skewness and Kurtosis:

Introduction

"Skewness essentially is a commonly used measure in descriptive statistics that characterizes the asymmetry of a data distribution, while kurtosis determines the heaviness of the distribution tails."

Understanding the shape of data is crucial while practicing data science. It helps to understand where the most information lies and analyze the outliers in a given data. In this article, we'll learn about the shape of data, the importance of skewness, and kurtosis in statistics. The types of ***skewness and kurtosis*** and Analyze the shape of data in the given dataset. Let's first understand what skewness and kurtosis is.

What Is Skewness?

Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.

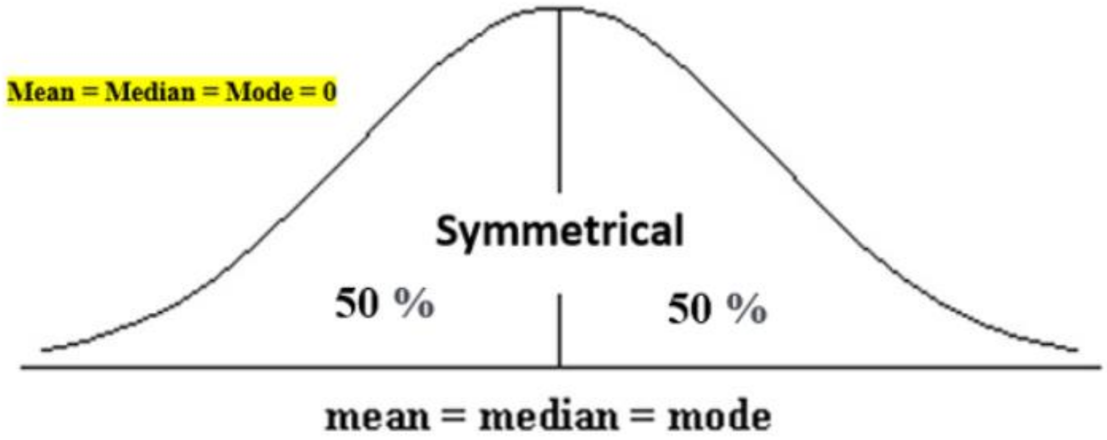
Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.

Depending on the model, skewness in the values of a specific independent variable (feature) may violate model assumptions or diminish the interpretation of feature importance.

A probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data exhibits skewness, which is a measure of asymmetry in statistics.

A skewed data set, typical values fall between the first quartile (Q1) and the third quartile (Q3).

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

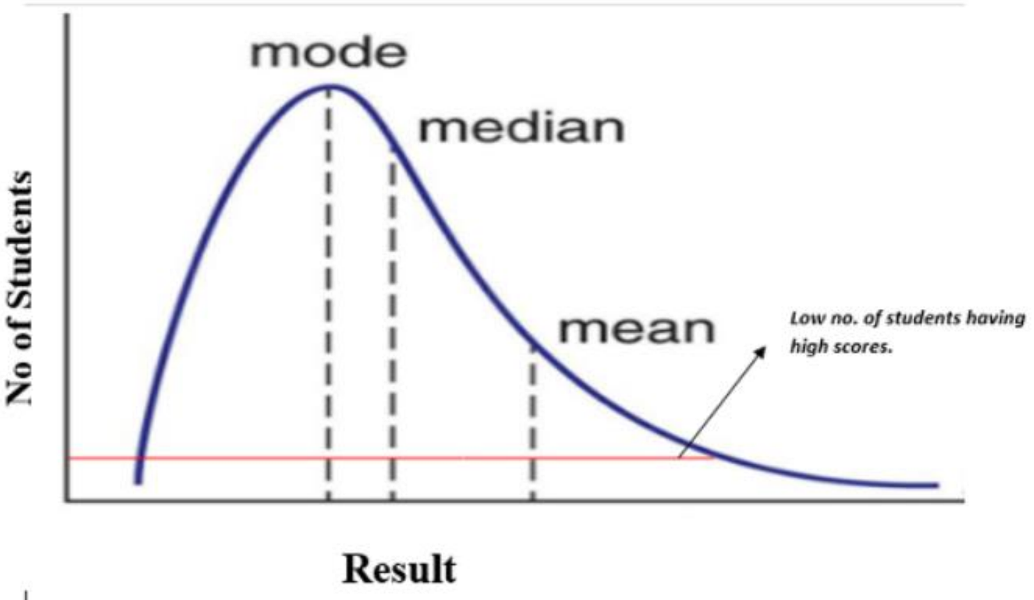


In a symmetrically distributed dataset, both the left-hand side and the right-hand side have an equal number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.

Types of Skewness

Positive Skewed or Right-Skewed (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero



Mean > Median > Mode

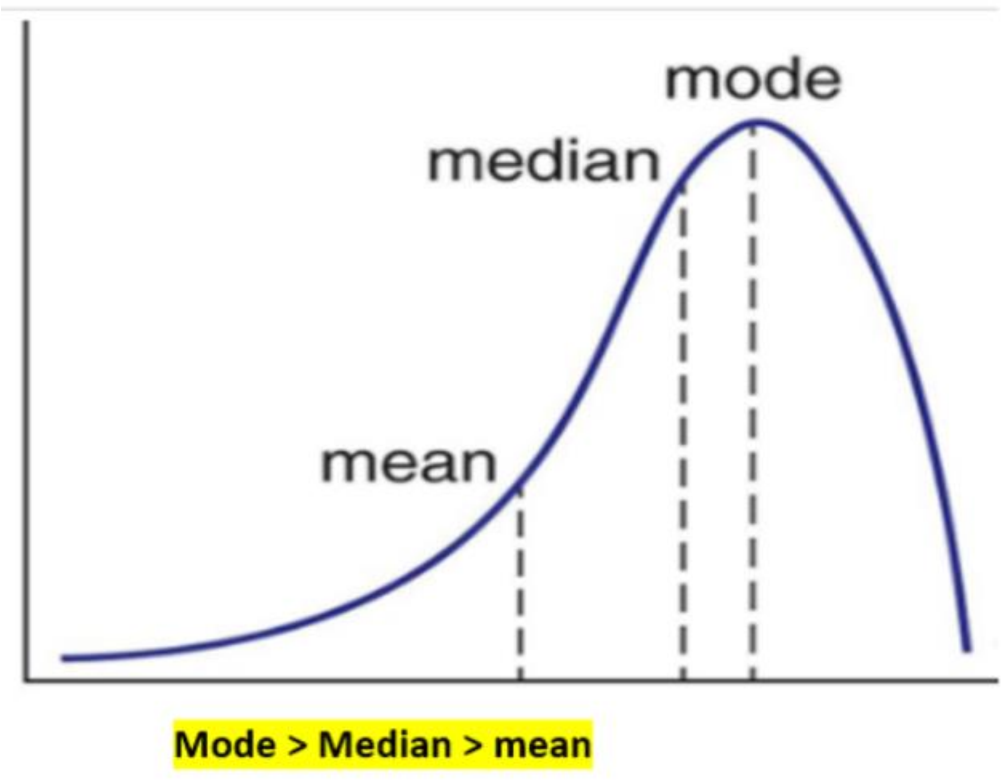
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.

Extreme positive skewness is not desirable for a distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

Negative Skewed or Left-Skewed (Negative Skewness)

A distribution with a long left tail, known as negatively skewed or left-skewed, stands in complete contrast to a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



Median is the middle value, and mode is the most frequent value. Due to an unbalanced distribution, the median will be higher than the mean.

How to Calculate the Skewness Coefficient?

Various methods can calculate skewness, with Pearson’s coefficient being the most commonly used method.

Pearson’s first coefficient of skewness

To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.

$$\textbf{Pearson's first coefficient} = \frac{\textbf{Mean} - \textbf{Mode}}{\textbf{Standard Deviation}}$$

As Pearson’s correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of **-1 to +1.** That accurately shows the range of the correlation values.

Pearson’s first coefficient of skewness is helping if the data present high mode. However, if the data exhibits low mode or multiple modes, it is preferable not to use Pearson’s first coefficient, and instead, Pearson’s second coefficient may be superior, as it does not depend on the mode.

Pearson's second coefficient of skewness

subtract the median from the *mean, multiply the difference by 3, and divide the product by the standard deviation.*

$$\textbf{Pearson's second coefficient} = \frac{\textbf{3 (Mean – Median)}}{\textbf{Standard Deviation}}$$

$$\textbf{Mean – Mode} \approx \textbf{3 (Mean – Median)}$$

Rule of thumb :

- **For skewness values between -0.5 and 0.5, the data exhibit approximate symmetry.**
- **Skewness values within the range of -1 and -0.5 (negative skewed) or 0.5 and 1(positive skewed) indicate slightly skewed data distributions.**
- **Data with skewness values less than -1 (negative skewed) or greater than 1 (positive skewed) are considered highly skewed.**

What Is Kurtosis?

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.

Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.

The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.

Peakedness in a data distribution is **the degree to which data values are concentrated around the mean**. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

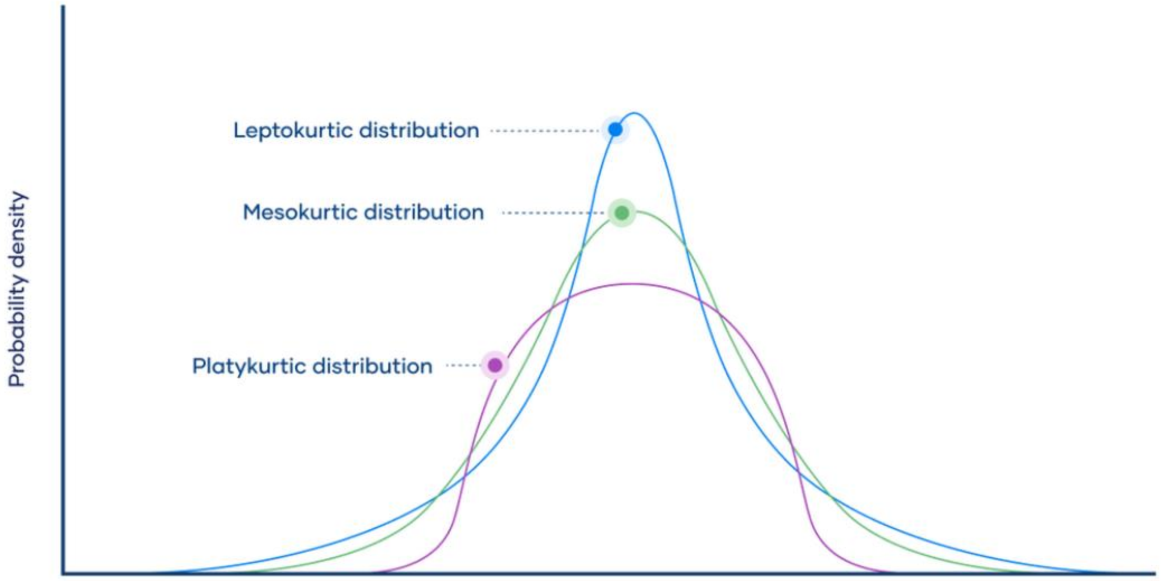
Types of Kurtosis

Kurtosis is a statistical measure that describes the shape of a probability distribution's tails relative to its peak. There are three main types of kurtosis:

1. **Mesokurtic**: A distribution with mesokurtic kurtosis has a similar peak and tail shape as the normal distribution. It has a kurtosis value of around 0, indicating that its tails are neither too heavy nor too light compared to a normal distribution.
2. **Leptokurtic**: A distribution with leptokurtic kurtosis has heavier tails and a sharper peak than the normal distribution. It has a positive kurtosis value, indicating that it has more extreme outliers than a normal distribution. This type of distribution is often associated with higher peakedness and a greater probability of extreme values.
3. **Platykurtic**: A distribution with platykurtic kurtosis has lighter tails and a flatter peak than the normal distribution. It has a negative kurtosis value, indicating that it has fewer extreme outliers than a normal distribution. This type of distribution is often associated with less peakedness and a lower probability of extreme values.

Types of Excess Kurtosis

1. *Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).*
2. *Mesokurtic (kurtosis same as the normal distribution).*
3. *Platykurtic or short-tailed distribution (kurtosis less than normal distribution).*



Leptokurtic *(Kurtosis > 3)*

Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

Platykurtic (Kurtosis < 3)

Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (Kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Difference Between Skewness and Kurtosis

Key Differences of Skewness and Kurtosis

1. Skewness evaluates how much a distribution deviates from symmetry, while Kurtosis gauges the degree of its peakiness or flatness.
2. Skewness is a measure derived from the third moment, whereas Kurtosis stems from the fourth moment.
3. The range of values for both Skewness and Kurtosis spans from negative infinity to positive infinity.
4. Perfect symmetry and normality are indicated by both zero skewness and zero kurtosis.
5. Skewness can impact the central tendency of a distribution, whereas kurtosis can influence its tail behavior.
6. Both Skewness and Kurtosis provide insights into the shape characteristics of distributions.

Session 11, 12, 13 & 14

Sampling

When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a **sample**. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.

The population can be defined in terms of geographical location, age, income, or many other characteristics.

It can be very broad or quite narrow: maybe you want to make inferences about the whole adult population of your country; maybe your research focuses on customers of a certain company, patients with a specific health condition, or students in a single school.

It is important to carefully define your target population according to the purpose and practicalities of your project.

If the population is very large, demographically mixed, and geographically dispersed, it might be difficult to gain access to a representative sample. A lack of a representative sample affects the validity of your results, and can lead to several research biases, particularly sampling bias.

Sampling frame

The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

Example

You are doing research on working conditions at a social media marketing company. Your population is all 1000 employees of the company. Your sampling frame is the company's HR database, which lists the names and contact details of every employee.

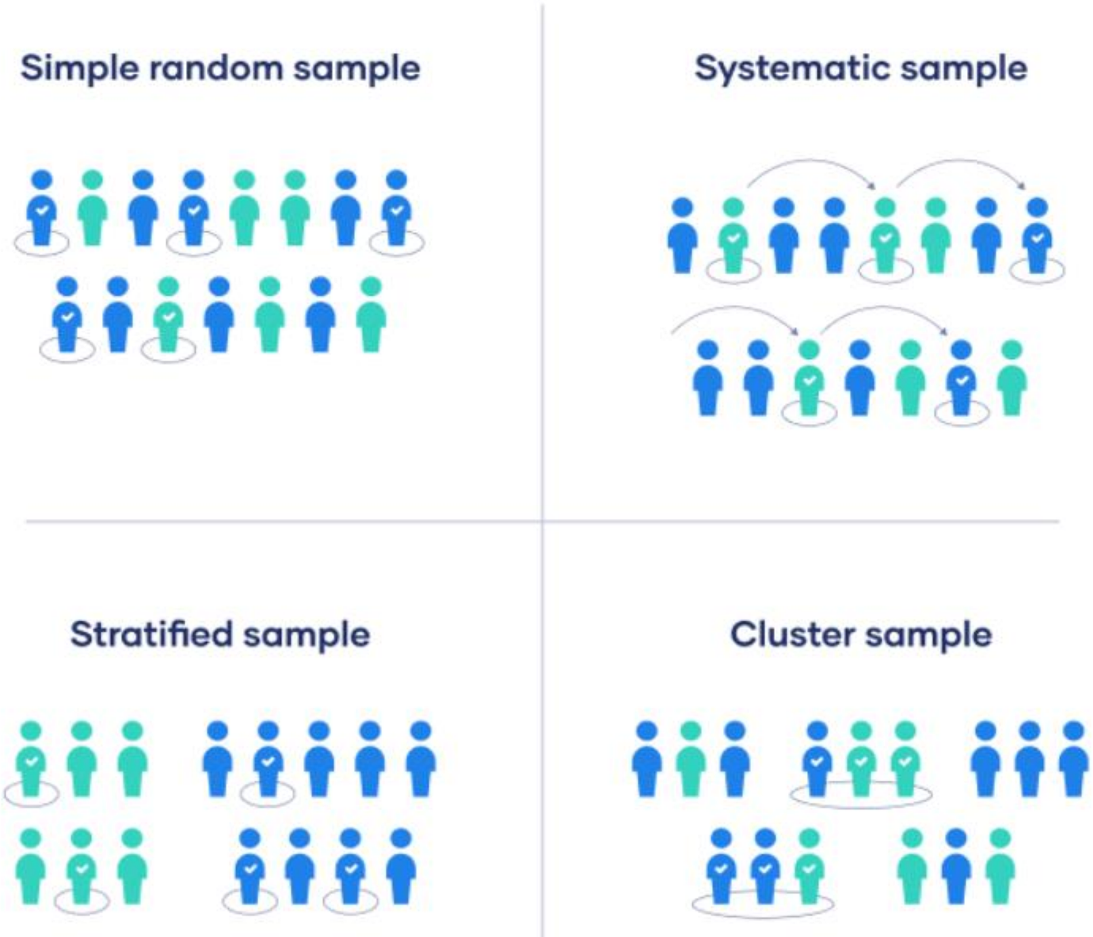
Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what you want to achieve with statistical analysis.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.



1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

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.

2. 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 samplingAll 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.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. 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).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling the company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. 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.

Non-probability sampling is a sampling method that uses non-random criteria like the availability, geographical proximity, or expert knowledge of the individuals you want to research in order to answer a research question.

Non-probability sampling is used when the population parameters are either unknown or not possible to individually identify. For example, visitors to a website that doesn't require users to create an account could form part of a non-probability sample.

Note that this type of sampling is at higher risk for research biases than probability sampling, particularly sampling bias.

Note

Be careful not to confuse probability and non-probability sampling.

- In **non-probability sampling**, each unit in your target population does **not** have an equal chance of being included. Here, you can form your sample using other considerations, such as convenience or a particular characteristic.
- In **probability sampling**, each unit in your target population **must** have an equal chance of selection.

Convenience sampling

Convenience sampling is primarily determined by convenience to the researcher.

This can include factors like:

- Ease of access
- Geographical proximity
- Existing contact within the population of interest

Convenience samples are sometimes called "accidental samples," because participants can be selected for the sample simply because they happen to be nearby when the researcher is conducting the data collection.

Example: Convenience samplingYou are investigating the association between daily weather and daily shopping patterns. To collect insight into people's shopping patterns, you decide to stand outside a major shopping mall in your area for a week, stopping people as they exit and asking them if they are willing to answer a few questions about their purchases.

Quota sampling

In **quota sampling**, you select a predetermined number or proportion of units, called a quota. Your quota should comprise subgroups with specific characteristics (e.g., individuals, cases, or organizations) and should be selected in a non-random manner.

Your subgroups, called **strata**, should be mutually exclusive. Your estimation can be based on previous studies or on other existing data, if there are any. This helps you determine how many units should be chosen from each subgroup. In the data collection phase, you continue to recruit units until you reach your quota.

TipYour respondents should be recruited non-randomly, with the end goal being that the proportions in each subgroup coincide with the estimated proportions in the population.

There are two types of quota sampling:

1. **Proportional** quota sampling is used when the size of the population is known. This allows you to determine the quota of individuals that you need to include in your sample in order to be representative of your population.

Example: Proportional quota samplingLet's say that in a certain company there are 1,000 employees. They are split into 2 groups: 600 people who drive to work, and 400 who take the train.
You decide to draw a sample of 100 employees. You would need to survey 60 drivers and 40 train-riders for your sample to reflect the proportion seen in the company.

2. **Non-proportional** quota sampling is used when the size of the population is unknown. Here, it's up to you to determine the quota of individuals that you are going to include in your sample in advance.

Example: Non-proportional quota samplingLet's say you are seeking opinions about the design choices on a website, but do not know how many people use it. You may decide to draw a sample of 100 people, including a quota of 50 people under 40 and a quota of 50 people over 40. This way, you get the perspective of both age groups.
Note that quota sampling may sound similar to stratified sampling, a probability sampling method where you divide your population into subgroups that share a common characteristic.

The key difference here is that in stratified sampling, you take a random sample from each subgroup, while in quota sampling, the sample selection is non-random, usually via convenience sampling. In other words, who is included in the sample is left up to the subjective judgment of the researcher.

Example: Quota samplingYou work for a market research company. You are seeking to interview 20 homeowners and 20 tenants between the ages of 45 and 60 living in a certain suburb.

You stand at a convenient location, such as a busy shopping street, and randomly select people to talk to who appear to satisfy the age criterion. Once you stop them, you must first determine whether they do indeed fit the criteria of belonging to the predetermined age range and owning or renting a property in the suburb.

Sampling continues until quotas for various subgroups have been selected. If contacted individuals are unwilling to participate or do not meet one of the conditions (e.g., they are over 60 or they do not live in the suburb), they are simply replaced by those who do. This approach really helps to mitigate nonresponse bias.

Self-selection (volunteer) sampling

Self-selection sampling (also called volunteer sampling) relies on participants who voluntarily agree to be part of your research. This is common for samples that need people who meet specific criteria, as is often the case for medical or psychological research.

In self-selection sampling, volunteers are usually invited to participate through advertisements asking those who meet the requirements to sign up. Volunteers are recruited until a predetermined sample size is reached.

Self-selection or volunteer sampling involves two steps:

1. Publicizing your need for subjects
2. Checking the suitability of each subject and either inviting or rejecting them

Example: Self-selection samplingSuppose that you want to set up an experiment to see if mindfulness exercises can increase the performance of long-distance runners. First, you need to recruit your participants. You can do so by placing posters near locations where people go running, such as parks or stadiums.
Your ad should follow ethical guidelines, making it clear what the study involves. It should also include more practical information, such as the types of participants required. In this case, you decide to focus on runners who can run at least 5 km and have no prior training or experience in mindfulness.
Keep in mind that not all people who apply will be eligible for your research. There is a high chance that many applicants will not fully read or understand what your study is about, or may possess disqualifying factors. It’s important to double-check eligibility carefully before inviting any volunteers to form part of your sample.

Snowball sampling

Snowball sampling is used when the population you want to research is hard to reach, or there is no existing database or other sampling frame to help you find them. Research about socially marginalized groups such as drug addicts, homeless people, or sex workers often uses snowball sampling.

To conduct a snowball sample, you start by finding one person who is willing to participate in your research. You then ask them to introduce you to others.

Alternatively, your research may involve finding people who use a certain product or have experience in the area you are interested in. In these cases, you can also use networks of people to gain access to your population of interest.

Example: Snowball samplingYou are studying homeless people living in your city. You start by attending a housing advocacy meeting, striking up a conversation with a homeless woman. You explain the purpose of your research and she agrees to participate. She invites you to a parking lot serving as temporary housing and offers to introduce you around.

In this way, the process of snowball sampling begins. You started by attending the meeting, where you met someone who could then put you in touch with others in the group.

When studying vulnerable populations, be sure to follow ethical considerations and guidelines.

Purposive (judgmental) sampling

Purposive sampling is a blanket term for several sampling techniques that choose participants deliberately due to qualities they possess. It is also called judgmental sampling, because it relies on the judgment of the researcher to select the units (e.g., people, cases, or organizations studied).

Purposive sampling is common in qualitative and mixed methods research designs, especially when considering specific issues with unique cases.

Note:

Unlike random samples—which deliberately include a diverse cross-section of ages, backgrounds, and cultures—the idea behind purposive sampling is to concentrate on people with particular characteristics, who will enable you to answer your research questions.

The sample being studied is not representative of the population, but for certain qualitative and mixed methods research designs, this is not an issue.

Hypothesis Testing

Hypothesis testing is a tool for making statistical inferences about the population data. It is an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy. Hypothesis testing provides a way to verify whether the results of an experiment are valid.

A null hypothesis and an alternative hypothesis are set up before performing the hypothesis testing. This helps to arrive at a conclusion regarding the sample obtained from the population. In this article, we will learn more about hypothesis testing, its types, steps to perform the testing, and associated examples.
Hypothesis testing uses sample data from the population to draw useful conclusions regarding the population probability distribution. It tests an assumption made about the data using different types of hypothesis testing methodologies. The hypothesis testing results in either rejecting or not rejecting the null hypothesis.

Hypothesis testing can be defined as a statistical tool that is used to identify if the results of an experiment are meaningful or not. It involves setting up a null hypothesis and an alternative hypothesis. These two hypotheses will always be mutually exclusive. This means that if the null hypothesis is true then the alternative hypothesis is false and vice versa. An example of hypothesis testing is setting up a test to check if a new medicine works on a disease in a more efficient manner.

Null Hypothesis

The null hypothesis is a concise mathematical statement that is used to indicate that there is no difference between two possibilities. In other words, there is no difference between certain characteristics of data. This hypothesis assumes that the outcomes of an experiment are based on chance alone. It is denoted as H_0. Hypothesis testing is used to conclude if the null hypothesis can be rejected or not. Suppose an experiment is conducted to check if girls are shorter than boys at the age of 5. The null hypothesis will say that they are the same height.

Alternative Hypothesis

The alternative hypothesis is an alternative to the null hypothesis. It is used to show that the observations of an experiment are due to some real effect. It indicates that there is a statistical significance between two possible outcomes and can be denoted as H_1 or H_a. For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.

Hypothesis Testing P Value

In hypothesis testing, the p value is used to indicate whether the results obtained after conducting a test are statistically significant or not. It also indicates the probability of making an error in rejecting or not rejecting the null hypothesis.This value is always a number between 0 and 1. The p value is compared to an alpha level, αα or significance level. The alpha level can be defined as the acceptable risk of incorrectly rejecting the null hypothesis. The alpha level is usually chosen between 1% to 5%.

Hypothesis Testing Critical region

All sets of values that lead to rejecting the null hypothesis lie in the critical region. Furthermore, the value that separates the critical region from the non-critical region is known as the critical value.

One Tailed Hypothesis Testing

One tailed hypothesis testing is done when the rejection region is only in one direction. It can also be known as directional hypothesis testing because the effects can be tested in one direction only. This type of testing is further classified into the right tailed test and left tailed test.

Right Tailed Hypothesis Testing

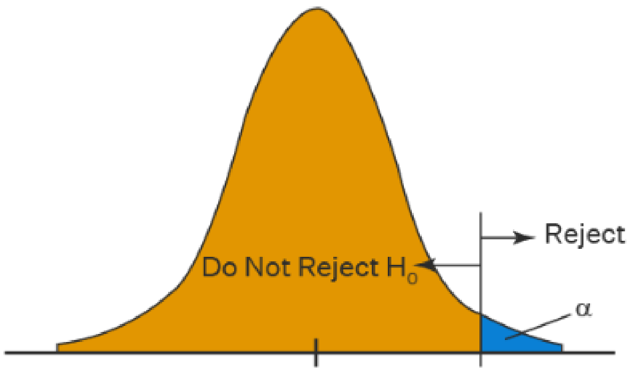
The right tail test is also known as the upper tail test. This test is used to check whether the population parameter is greater than some value. The null and alternative hypotheses for this test are given as follows:

H_0: The population parameter = some value

H_1: The population parameter is > some value.

If the test statistic has a greater value than the critical value then the null hypothesis is rejected

Right Tail Hypothesis Testing



Left Tailed Hypothesis Testing

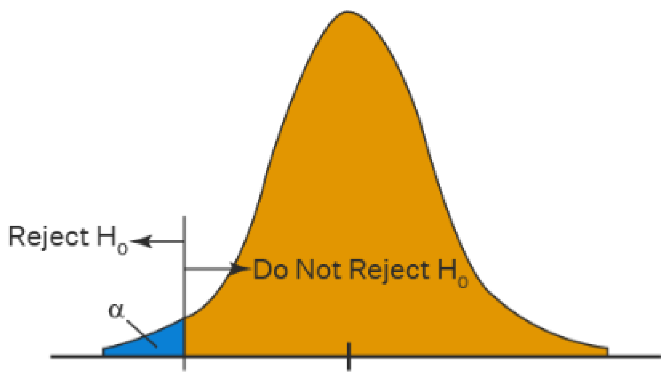
The left tail test is also known as the lower tail test. It is used to check whether the population parameter is less than some value. The hypotheses for this hypothesis testing can be written as follows:

H_0: The population parameter is= some value

H_1: The population parameter is < some value.

The null hypothesis is rejected if the test statistic has a value lesser than the critical value.

Left Tail Hypothesis Testing



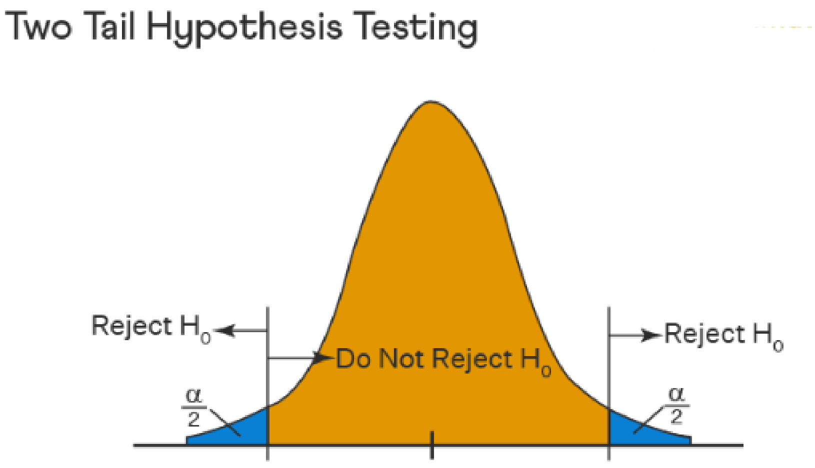
Two Tailed Hypothesis Testing

In this hypothesis testing method, the critical region lies on both sides of the sampling distribution. It is also known as a non - directional hypothesis testing method. The two-tailed test is used when it needs to be determined if the population parameter is assumed to be different than some value. The hypotheses can be set up as follows:

H0: the population parameter = some value

H1: the population parameter ≠ some value

The null hypothesis is rejected if the test statistic has a value that is not equal to the critical value.



Types of errors for a Hypothesis Test

The goal of any hypothesis testing is to make a decision. In particular, we will decide whether to reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_1. Although we would like always to be able to make a correct decision, we must remember that the decision will be based on sample information, and thus we are subject to make one of two types of error, as defined in the accompanying boxes. A Type I error is the error of rejecting the null hypothesis when it is true. The probability of committing a Type I error is usually denoted by α. A Type II error is the error of accepting the null hypothesis when it is false. The probability of making a Type II error is usually denoted by β. The null hypothesis can be either true or false and based on the sample drawn we make a conclusion either to reject or not to reject the null hypothesis. Thus, there are four possible situations that may arise in testing a hypothesis

		True "State of Nature"	
		Null Hypothesis	*Alternative Hypothesis*
Conclusions	*Do not reject Null Hypothesis*	No Error	Type II error
	Reject Null Hypothesis	Type I error	No Error

Hypothesis Testing Steps

Hypothesis testing can be easily performed in five simple steps. The most important step is to correctly set up the hypotheses and identify the right method for hypothesis testing. The basic steps to perform hypothesis testing are as follows:

- Step 1: Set up the null hypothesis by correctly identifying whether it is the left-tailed, right-tailed, or two-tailed hypothesis testing.
- Step 2: Set up the alternative hypothesis.
- Step 3: Choose the correct significance level, α, and find the critical value.
- Step 4: Calculate the correct test statistic and p-value.
- Step 5: Compare the test statistic with the critical value or compare the p-value with αα to arrive at a conclusion. In other words, decide if the null hypothesis is to be rejected or not.

Hypothesis Testing Example

The best way to solve a problem on hypothesis testing is by applying the 5 steps mentioned in the previous section. Suppose a researcher claims that the mean average weight of men is greater than 100kgs with a standard deviation of 15kgs. 30 men are chosen with an average weight of 112.5 Kgs. Using hypothesis testing, check if there is enough evidence to support the researcher's claim. The confidence interval is given as 95%.

Step 1: This is an example of a right-tailed test. Set up the null hypothesis as H0H0: μμ = 100.

Step 2: The alternative hypothesis is given by H_1: $\mu > 100$.

Step 3: As this is a one-tailed test, $\alpha = 100\% - 95\% = 5\%$. This can be used to determine the critical value. $1 - \alpha = 1 - 0.05 = 0.95$

0.95 gives the required area under the curve. Now using a normal distribution table, the area 0.95 is at z = 1.645. A similar process can be followed for a t-test. The only additional requirement is to calculate the degrees of freedom given by n - 1.

Step 4: Calculate the z test statistic. This is because the sample size is 30. Furthermore, the sample and population means are known along with the standard deviation.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\mu = 100, \bar{x} = 112.5, n = 30, \sigma = 15$$

$$z = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}} = 4.56$$

Step 5: Conclusion. As 4.56 > 1.645 thus, the null hypothesis can be rejected.

Z Test

Z test is a statistical test that is conducted on data that approximately follows a normal distribution. The z test can be performed on one sample, two samples, or on proportions for hypothesis testing. It checks if the means of two large samples are different or not when the population variance is known.

A z test can further be classified into left-tailed, right-tailed, and two-tailed hypothesis tests depending upon the parameters of the data. In this article, we will learn more about the z test, its formula, the z test statistic, and how to perform the test for different types of data using examples.

What is Z Test?

A z test is a test that is used to check if the means of two populations are different or not provided the data follows a normal distribution. For this purpose, the null hypothesis and the alternative hypothesis must be set up and the value of the z test statistic must be calculated. The decision criterion is based on the z critical value.

Z Test Definition

A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known. The null hypothesis of a z test can be rejected if the z test statistic is statistically significant when compared with the critical value.

Z Test Formula

The z test formula compares the z statistic with the z critical value to test whether there is a difference in the means of two populations. In hypothesis testing, the z critical value divides the distribution graph into the acceptance and the rejection regions. If the test statistic falls in the rejection region then the null hypothesis can be rejected otherwise it cannot be rejected. The z test formula to set up the required hypothesis tests for a one sample and a two-sample z test are given below.

One-Sample Z Test

A one-sample z test is used to check if there is a difference between the sample mean and the population mean when the population standard deviation is known. The formula for the z test statistic is given as follows:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation and n is the sample size.

The algorithm to set a one sample z test based on the z test statistic is given as follows:

Left Tailed Test:

Null Hypothesis: $H_0 : \mu=\mu_0$

Alternate Hypothesis: $H_1 : \mu<\mu_0$

Decision Criteria: If the z statistic < z critical value then reject the null hypothesis.

Right Tailed Test:

Null Hypothesis: $H_0 : \mu=\mu_0$

Alternate Hypothesis: $H_1 : \mu>\mu_0$

Decision Criteria: If the z statistic > z critical value then reject the null hypothesis.

Two Tailed Test:

Null Hypothesis: $H_0 : \mu=\mu_0$

Alternate Hypothesis: $H_1 : \mu\neq\mu_0$

Decision Criteria: If the z statistic > z critical value then reject the null hypothesis.

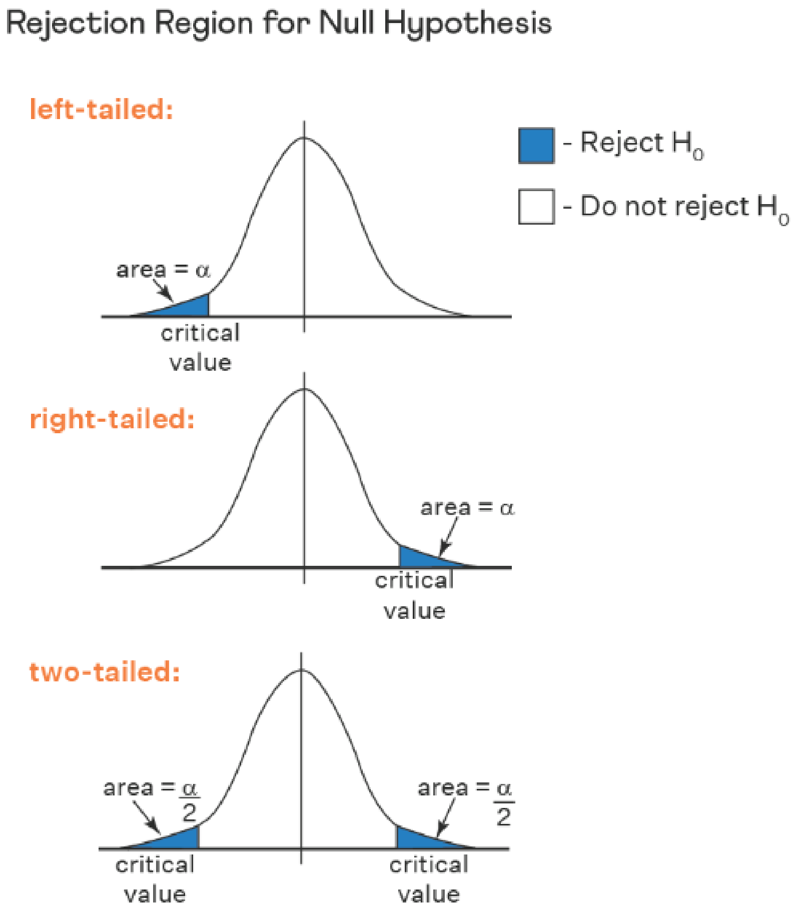
Two Sample Z Test

A two-sample z test is used to check if there is a difference between the means of two samples. The z test statistic formula is given as follows:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\overline{x_1}$, μ_1, σ^2_1 are the sample mean, population mean and population variance respectively for the first sample. $\overline{x_2}$, μ_2, σ^2_2 are the sample mean, population mean and population variance respectively for the second sample.

The two-sample z test can be set up in the same way as the one-sample test. However, this test will be used to compare the means of the two samples. For example, the null hypothesis is given as $H_0 : \mu_1 = \mu_2$



Z Test for Proportions

A z test for proportions is used to check the difference in proportions. A z test can either be used for one proportion or two proportions. The formulas are given as follows.

One Proportion Z Test

A one proportion z test is used when there are two groups and compares the value of an observed proportion to a theoretical one. The z test statistic for a one proportion z test is given as follows:

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Here, p is the observed value of the proportion, p_0 is the theoretical proportion value and n is the sample size.

The null hypothesis is that the two proportions are the same while the alternative hypothesis is that they are not the same.

Two Proportion Z Test

A two-proportion z test is conducted on two proportions to check if they are the same or not. The test statistic formula is given as follows:

$$z = \frac{p_1 - p_2 - 0}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$\text{where } p = \frac{x_1+x_2}{n_1+n_2}$$

p_1 is the proportion of sample 1 with sample size n_1 and x1 number of trials.

p_2 is the proportion of sample 2 with sample size n_2 and x2 number of trials.

- Z test is a statistical test that is conducted on normally distributed data to check if there is a difference in means of two data sets.
- The sample size should be greater than 30 and the population variance must be known to perform a z test.
- The one-sample z test checks if there is a difference in the sample and population mean,
- The two sample z test checks if the means of two different groups are equal.

t-test Formula

The t-test formula helps us to compare the average values of two data sets and determine if they belong to the same population or are they different. The t-score is compared with the critical value obtained from the t-table. The large t-score indicates that the groups are different and a small t-score indicates that the groups are similar.

What Is the T-test Formula?

The t-test formula is applied to the sample population. The t-test formula depends on the mean, variance, and standard deviation of the data being compared. There are 3 types of t-tests that could be performed on the n number of samples collected.

- One-sample test,
- Independent sample t-test and
- Paired samples t-test

The critical value is obtained from the t-table looking for the degree of freedom (df = n-1) and the corresponding α value (usually 0.05 or 0.1). If the t-test obtained statistically > CV then the initial hypothesis is wrong and we conclude that the results are significantly different.

One-Sample T-Test Formula

For comparing the mean of a population $\overline{x}$ from n samples, with a specified theoretical mean μ, we use a one-sample t-test.

$$t = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where $\sigma/\sqrt{n}$ is the standard error

t-Test Formula

$$t = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where,
'x' bar is the mean of the sample,
μ is the assumed mean,
σ is the standard deviation
and n is the number of observations

Independent Sample T-Test

Students t-test is used to compare the mean of two groups of samples. It helps evaluate if the means of the two sets of data are statistically significantly different from each other.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

T-test Formula

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$df = n_1 + n_2 - 1$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}{n_1 - 1}$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2} (x_j - \bar{x}_2)^2}{n_2 - 1}$$

where

- t = Student's t-test
- x_1 = mean of first group
- x_2= mean of second group
- s_1 = standard deviation of group 1
- s_2 = standard deviation of group 1
- n_1= number of observations in group 1
- n_2= number of observations in group 2

Paired Samples T-Test

Whenever two distributions of the variables are highly correlated, they could be pre and post test results from the same people. In such cases, we use the paired samples t-test.

$$t = \frac{\Sigma(x_1 - x_2)}{\frac{s}{\sqrt{n}}}$$

where

t = Student's t-test

$x_1 - x_2$ = Difference mean of the pairs

s= standard deviation

n = sample size

Z Test	T-Test
A z test is a statistical test that is used to check if the means of two data sets are different when the population variance is known.	A t-test is used to check if the means of two data sets are different when the population variance is not known.
The sample size is greater than or equal to 30.	The sample size is lesser than 30.
The data follows a normal distribution.	The data follows a student-t distribution.
The one-sample z test statistic is given by $\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$	The t test statistic is given as $\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ where s is the sample standard deviation

ANOVA

ANOVA is used to analyze the differences among the means of various groups using certain estimation procedures. ANOVA means analysis of variance. ANOVA is a statistical significance test that is used to check whether the null hypothesis can be rejected or not during hypothesis testing.

An ANOVA can be either one-way or two-way depending upon the number of independent variables. In this article, we will learn more about an ANOVA test, the one-way ANOVA and two-way ANOVA, its formulas and see certain associated examples.

What is ANOVA?

ANOVA, in its simplest form, is used to check whether the means of three or more populations are equal or not. The ANOVA applies when there are more than two independent groups. The goal of the ANOVA is to check for variability within the groups as well as the variability among the groups. The ANOVA statistic is given by the f test.

ANOVA Definition

ANOVA can be defined as a type of test used in hypothesis testing to compare whether the means of two or more groups are equal or not. This test is used to check if the null hypothesis can be rejected or not depending upon the statistical significance exhibited by the parameters. The decision is made by comparing the ANOVA statistic with the critical value.

ANOVA Example

Suppose it needs to be determined if consumption of a certain type of tea will result in a mean weight loss. Let there be three groups using three types of tea - green tea, earl grey tea, and jasmine tea. Thus, to compare if there was any mean weight loss exhibited by a certain group, the ANOVA (one way) will be used.

Suppose a survey was conducted to check if there is an interaction between income and gender with anxiety level at job interviews. To conduct such a test a two-way ANOVA will be used.

ANOVA Formula

ANOVA Test Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

ANOVA Table

The ANOVA formulas can be arranged systematically in the form of a table. This ANOVA table can be summarized as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

Chi Square Formula

Chi-square formula is used to compare two or more statistical data sets. The chi-square formula is used in data that consist of variables distributed across various categories and helps us to know whether that distribution is different from what one would expect by chance.

Example: You research two groups of women and put them in categories of student, employed or self-employed.

	Group 1	Group 2
Student	40	30
Employed	89	67
Self-employed	3	7

The numbers collected are different, but you now want to know

- Is that just a random occurrence? Or
- Is there any correlation?

What is the Chi Square Formula?

The chi-squared test checks the difference between the observed value and the expected value. Chi-Square shows or in a way check the relationship between two categorical variables which can be can be calculated by using the given observed frequency and expected frequency.

Chi Square Formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where

O_i = observed value (actual value)

E_i = expected value

Applications of Chi Square Formula

- used by Biologists to determine if there is a significant association between the two variables, such as the association between two species in a community.
- used by Genetic analysts to interpret the numbers in various phenotypic classes.
- used in various statistical procedures to help to decide if to hold onto or reject the hypothesis.
- used in the medical literature to compare the incidence of the same characteristics in two or more groups.

Non-Parametric Test

Non-parametric test is a statistical analysis method that does not assume the population data belongs to some prescribed distribution which is determined by some parameters. Due to this, a non-parametric test is also known as a distribution-free test. These tests are usually based on distributions that have unspecified parameters.

A non-parametric test acts as an alternative to a parametric test for mathematical models where the nature of parameters is flexible. Usually, when the assumptions of parametric tests are violated then non-parametric tests are used. In this article, we will learn more about a non-parametric test, the types, examples, advantages, and disadvantages.

What is Non-Parametric Test in Statistics?

A non-parametric test in statistics does not assume that the data has been taken from a normal distribution. A normal distribution belongs to a parametrized family of probability distributions and includes parameters such as mean, variance, standard deviation, etc. Thus, a non-parametric test does not make assumptions about the probability distribution's parameters.

Non-Parametric Test Definition

A non-parametric test can be defined as a test that is used in statistical analysis when the data under consideration does not belong to a parametrized family of distributions. When the data does not meet the requirements to perform a parametric test, a non-parametric test is used to analyze it.

Reasons to Use Non-Parametric Tests

It is important to access when to apply parametric and non-parametric tests in order to arrive at the correct statistical inference. The reasons to use a non-parametric test are given below:

- When the distribution is skewed, a non-parametric test is used. For skewed distributions, the mean is not the best measure of central tendency, hence, parametric tests cannot be used.
- If the size of the data is too small then validating the distribution of the data becomes difficult. Thus, in such cases, a non-parametric test is used to analyze the data.
- If the data is nominal or ordinal, a non-parametric test is used. This is because a parametric test can only be used for continuous data.

Mann-Whitney U Test

This non-parametric test is analogous to t-tests for independent samples. To conduct such a test the distribution must contain ordinal data. It is also known as the Wilcoxon rank sum test.

Null Hypothesis: H_0: The two populations under consideration must be equal.

Test Statistic: U should be smaller of

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ or $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

where, R_1 is the sum of ranks in group 1 and R_2 is the sum of ranks in group 2.

Decision Criteria: Reject the null hypothesis if $U \leq$ critical value.