Piyush Wairale

GATE

Data Science & AI
2025

Probability & Statistics

For GATE DA Course & Test Series

Visit:

www.piyushwairale.com

GATE Data Science and AI Study Materials

Probability & Statistics

by Piyush Wairale

Instructions:

- Kindly go through the lectures/videos on our website www.piyushwairale.com
- Read this study material carefully and make your own handwritten short notes. (Short notes must not be more than 5-6 pages)
- Attempt the mock tests available on portal.
- Revise this material at least 5 times and once you have prepared your short notes, then revise your short notes twice a week
- If you are not able to understand any topic or required a detailed explanation and if there are any typos or mistake in study materials. Mail me at piyushwairale100@gmail.com

Contents

1 Fundamental Principles of Counting 5
1.1 Addition Principle 5
1.2 Multiplication Principle 5

2 Permutations 5

3 Combinations 6

4 Introduction to Probability 7
4.1 Definition of Probability 7
4.2 Theorems of Probability 8

5 Conditional Probability and Bayes Theorem 9
5.1 Bayes' Theorem 10

6 Discrete and Continuous Random Variables 12
6.1 Probability Mass Function (PMF) 12
6.2 Probability Density Function (PDF) 12

7 Expectation (Mean), Variance, and Standard Deviation 13
7.1 Expectation (Mean) 13
7.2 Variance 14
7.3 Standard Deviation 15
7.4 Covariance and Correlation 15

8 Discrete Distributions 17
8.1 Discrete Uniform Distribution 17
8.2 Binomial Distribution 17
8.3 Poisson Distribution 17

9 Continuous Distributions 19
9.1 Uniform/Rectangular Distribution 19
9.2 Normal Distribution 19
9.3 Standard Normal Distribution 20
9.4 Exponential Distribution 21

10 Joint Distribution of Random Variables 22
10.1 Joint Probability Density Function 22
10.2 Conditional Probability Functions of Random Variables 23
10.3 Independent Random Variables 23

11 Mean, Median, Mode and Standard Deviation 24
11.1 Mean 24
11.2 Median 24
11.3 Mode 24
11.4 Standard Deviation 25

12 t-distribution 26
12.1 Key Characteristics: 26
    12.1.1 Comparing the t-distribution and Normal Distribution: 26
12.2 Formula for the t-distribution: 26
12.3 Mean, Median, Mode and SD 27

13 Chi-Square Distribution 28
13.1 Key Characteristics of the Chi-Square Distribution: 28
13.2 Key Formulas for the Chi-Square Tests: 29

14 Hypothesis Testing **30**
14.1 Key Terms: 30
14.2 Steps in Hypothesis Testing: 31
14.3 Significance Level and Confidence Level in Terms of t-Value 32

15 t-test **34**

16 z-test **36**

17 chi-test **37**

18 Central Limit Theorem **40**
18.1 Key Points of the Central Limit Theorem: 40

INSTAGRAM
FOLLOW US!
PIYUSH WAIRALE
INSTAGRAM
Linkedn in
FACEGRAM
FACEBOOK
FACEBROOK
Lintegok
Youedin
LINKEDIN
YouTube
LINKEDLIN
TELEGRAM
FOLLOW US !!

LinkedIn

Youtube Channel

Instagram

Telegram Group

Facebook

Download Andriod App

1 Fundamental Principles of Counting

1.1 Addition Principle

If there are n_1 ways to perform task 1 and n_2 ways to perform task 2, and the tasks cannot be done simultaneously, then there are $n_1 + n_2$ ways to perform either task.

Example: If there are 3 routes from city A to B and 2 routes from city A to C, and you want to go to either B or C, there are $3 + 2 = 5$ ways.

1.2 Multiplication Principle

If there are n_1 ways to perform task 1 and n_2 ways to perform task 2, then there are $n_1 \times n_2$ ways to perform both tasks in succession.

Example: If you have 3 shirts and 2 pants, there are $3 \times 2 = 6$ ways to choose an outfit.

2 Permutations

A **permutation** is an arrangement of objects in a specific order. The order of arrangement is important.

Note: In a permutation, the order of arrangement of the objects is important. Thus abc is a different permutation from bca.
In a combination, the order in which objects are selected does not matter.
Thus abc and bca are the same combination.

Formula for Permutations

- **Number of permutations of n distinct objects taken r at a time:**

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

  where $n!$ (n factorial) is the product of all positive integers up to n.

Permutations of All Objects

- **Number of permutations of n distinct objects:**

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

Permutations with Repetition

- **Number of permutations of n objects where n_1 are of type 1, n_2 are of type 2, ..., n_k are of type k:**

$$\frac{n!}{n_1! \times n_2! \times \cdots \times n_k!}$$

Circular Permutations

- **Number of circular permutations of n distinct objects:**
  - Without distinguishing rotations:

$$(n-1)!$$

  - If reflections are considered identical:

$$\frac{(n-1)!}{2}$$

Permutations with Restrictions

- **Derangements:** Number of permutations where no element appears in its original position.
  - **Number of derangements of n objects (denoted by $!n$):**

$$!n = n!\sum_{k=0}^{n}\frac{(-1)^k}{k!}$$

  - **Approximate formula:**

$$!n \approx \frac{n!}{e}$$

where e is the base of the natural logarithm ($e \approx 2.71828$).

3 Combinations

A **combination** is a selection of objects where the order does not matter.

Formula for Combinations

- **Number of combinations of n distinct objects taken r at a time:**

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

Properties of Combinations

- **Symmetry Property:**

$$nCr = nC(n-r)$$

- **Addition Formula:**

$$nCr + nC(r-1) = (n+1)Cr$$

Combinations with Repetition

- **Number of ways to choose r objects from n types with unlimited repetitions:**

$$C(n+r-1,r) = \frac{(n+r-1)!}{r! \times (n-1)!}$$

Pascal's Triangle and Binomial Theorem

Pascal's Triangle

- A triangular array where each number is the sum of the two numbers directly above it.
- **The n-th row corresponds to the coefficients in the expansion of $(a+b)^n$.**

Binomial Theorem

- **Expansion of $(a+b)^n$:**

$$(a+b)^n = \sum_{r=0}^{n} nCr\, a^{n-r} b^r$$

Applications:

- Finding coefficients in binomial expansions.
- Solving combinatorial identities.

4 Introduction to Probability

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, such an action is called a *Random Experiment*. One performance of such an experiment is called a *Trial*.

Sample Space

The set of all possible outcomes of a random experiment is called the sample space. All the elements of the sample space together are called as *exhaustive cases*. The number of elements of the sample space i.e. the number of exhaustive cases is denoted by $n(S)$ or N or n.

Event

Any subset of the sample space is called as an *Event* and is denoted by some capital letter like A, B, C, or A_1, A_2, A_3, ..., or B_1, B_2, ... etc.

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 $n(A)$ or N_A or n_A.

Mutually Exclusive Events or Disjoint Events

Two events A and B are said to be mutually exclusive or disjoint if $A \cap B = \emptyset$, i.e. if there is no element common to A and B.

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.

4.1 Definition of Probability

Consider a random experiment which results in a sample space containing $n(S)$ cases which are exhaustive, mutually exclusive, and equally likely. Suppose, out of $n(S)$ cases, $n(A)$ cases are favourable to an event A. Then, the probability of event A is denoted by $P(A)$ and is defined as follows:

$$P(A) = \frac{n(A)}{n(S)}$$

Complement of an Event

The complement of an event A is denoted by A' and it contains all the elements of the sample space which do not belong to A.

For example: Random experiment: an unbiased die is rolled.

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

$$A = \{1, 4\}, \quad A' = \{2, 3, 5, 6\}$$

$$P(A) + P(A') = 1 \quad \text{i.e.} \quad P(A) = 1 - P(A')$$

For any events A and B,

$$P(A \cap B) = P(A) \cdot P(B)$$

Independent Events

Two events A and B are said to be independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

Note: If A and B are independent, then

- A and B are independent
- A' and B are independent
- A and B' are independent
- A' and B' are independent

4.2 Theorems of Probability

Addition Theorem

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 i.e. at least one of A B
2. $A' \cap B'$: neither A nor B i.e. none of A B
3. $A \cup B$ and $A' \cap B'$ are complement to each other
$$P(A' \cup B') = 1 - P(A \cup B)$$
4. If A and B are mutually exclusive, $P(A \cap B) = 0$
$$P(A \cup B) = P(A) + P(B)$$
5. For three events, we have:
$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

5 Conditional Probability and Bayes Theorem

The **conditional probability** of an event A given that another event B has occurred is denoted by $P(A|B)$ and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0.$$

Conditional probability measures the probability of event A occurring when it is known that event B has already occurred.

Properties

1. **Multiplication Rule:**
$$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A).$$
2. **If A and B are independent:**
$$P(A|B) = P(A).$$
3. **Total Probability:**
$$\sum_i P(A_i|B) = 1,$$
where $\{A_i\}$ are mutually exclusive and exhaustive events.

Example

Suppose we have a standard deck of 52 playing cards. What is the probability that a card drawn at random is a King given that it is a face card?

Solution:

- Total face cards: Jack, Queen, King in each suit ⇒ 12 cards.
- Number of Kings: 4.
- Conditional probability:

$$P(\text{King}|\text{Face card}) = \frac{P(\text{King and Face card})}{P(\text{Face card})} = \frac{\frac{4}{52}}{\frac{12}{52}} = \frac{4}{12} = \frac{1}{3}.$$

Law of Total Probability

If $\{B_1, B_2, \ldots, B_n\}$ is a partition of the sample space S (i.e., mutually exclusive and exhaustive events), then for any event A:

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

The law of total probability allows us to compute the probability of an event A by considering all the different ways A can occur through the events B_i.

Example

A factory has three machines M_1, M_2, and M_3 producing 30%, 45%, and 25% of the total products, respectively. The defect rates are 1%, 2%, and 3%, respectively. What is the probability that a randomly selected product is defective?

Solution:

- Let D be the event that a product is defective.
- Using the law of total probability:

$$P(D) = P(M_1)P(D|M_1) + P(M_2)P(D|M_2) + P(M_3)P(D|M_3)$$

$$P(D) = 0.30 \times 0.01 + 0.45 \times 0.02 + 0.25 \times 0.03 = 0.0175.$$

5.1 Bayes' Theorem

Bayes' theorem relates the conditional and marginal probabilities of random events. For events A and B with $P(B) > 0$:

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

Extended Form (Multiple Events)

For a partition $\{B_1, B_2, \ldots, B_n\}$ of the sample space S:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Bayes' theorem allows us to update the probability of an event based on new evidence or information.

Example

Using the previous factory example, if a randomly selected product is found to be defective, what is the probability it was produced by machine M_2?

Solution:

- We need to find $P(M_2|D)$.
- Using Bayes' theorem:

$$P(M_2|D) = \frac{P(D|M_2)P(M_2)}{P(D)}$$

$$P(M_2|D) = \frac{0.02 \times 0.45}{0.0175} = \frac{0.009}{0.0175} = \frac{9}{17.5} = \frac{18}{35} \approx 0.5143.$$

Independent Events

Events A and B are independent if:

$$P(A \cap B) = P(A) \times P(B).$$

Conditional Independence

Events A and B are conditionally independent given event C if:

$$P(A \cap B|C) = P(A|C) \times P(B|C).$$

Relation to Conditional Probability

If A and B are independent, then:

$$P(A|B) = P(A), \quad P(B|A) = P(B).$$

Total Probability and Bayes' Theorem in Continuous Cases

Continuous Random Variables

For continuous random variables, the concepts extend using probability density functions (pdfs).

Law of Total Probability

If X is a continuous random variable with pdf $f_X(x)$, and Y is another variable, then:

$$P(Y) = \int_{-\infty}^{\infty} P(Y|X = x) f_X(x)\, dx.$$

Bayes' Theorem

Bayes' theorem for continuous variables:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

6 Discrete and Continuous Random Variables

A **discrete random variable** takes the values that are finite or countable. For example, when we consider the experiment of tossing 3 coins, the number of heads can be appreciated as a discrete random variable X. X would take $0, 1, 2, 3$ as possible values.

A **continuous random variable** takes values in the form of intervals. Also, in the case of a continuous random variable $P(X = c) = 0$, where c is a specified point. Heights and weights of people, area of land held by individuals, etc., are examples of continuous random variables.

6.1 Probability Mass Function (PMF)

If X is a discrete random variable, which can take the values $x_1, x_2, \ldots$ and $f(x)$ denote the probability that X takes the value x_r, then $p(x)$ is called the Probability Mass Function (PMF) of X.

$$p(x_r) = P(X = x_r)$$

The values that X can take and the corresponding probabilities determine the probability distribution of X. We also have the following conditions:

1. $p(x) \geq 0$
2. $\sum p(x) = 1$

6.2 Probability Density Function (PDF)

If X is a continuous random variable, then a function $f(x)$, $x \in I$ (interval), is called a Probability Density Function. The probability statements are made as $P(X \in I) = \int_I f(x)\,dx$.

We also have:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

The probability $P(X \leq x)$ is called the cumulative distribution function (CDF) of X and is denoted by $F(X)$. It is a point function. It is defined for discrete and continuous random variables. The following are the properties of probability distribution function $F(x)$:

1. $F(x) \geq 0$
2. $F(x)$ is non-decreasing, i.e., for $x > y$, $F(x) \geq F(y)$
3. $F(x)$ is right continuous
4. $F(-\infty) = 0$ and $F(+\infty) = 1$
5. $P(a < X \leq b) = F(b) - F(a)$

For a continuous random variable:

- $Pr\{x \leq X \leq x + dx\} = F(x + dx) - F(x) = f(x)\,dx$ where dx is very small
- $f(x) = \frac{d}{dx}[F(x)]$ where:
  1. $f(x) \geq 0$ for all $x \in R$
  2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

7 Expectation (Mean), Variance, and Standard Deviation

7.1 Expectation (Mean)

Mathematical Expectation is the weighted mean of values of a variable.

If X is a random variable which can assume any one of the values $x_1, x_2, \ldots, x_n$, with the respective probabilities $p_1, p_2, \ldots, p_n$, then the mathematical expectation of X is given by

$$E(X) = p_1x_1 + p_2x_2 + \cdots + p_nx_n$$

Expected value for a discrete random variable is given by:

$$E(X) = \sum_{i=1}^{n} x_i P(x_i)$$

For a continuous random variable,

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{where} \quad f(x) \text{ is the PDF of } X.$$

Expected value for a continuous random variable is given by:

$$E(X) = \int_{a}^{b} x f(x)\, dx$$

Properties of Expectation

Linearity of Expectation

1. **Addition:**

$$E[X + Y] = E[X] + E[Y].$$

2. **Scalar Multiplication:**

$$E[aX] = a\,E[X],$$

   where a is a constant.

3. **General Linearity:**

$$E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E[X_i],$$

   where a_i are constants and X_i are random variables.

Note: The linearity property holds regardless of whether the random variables are independent or not.

Expectation of a Constant

$$E[c] = c,$$

where c is a constant.

Expectation of a Function of a Random Variable

For a function $g(X)$:

- **Discrete:**

$$E[g(X)] = \sum_{x} g(x)\, P(X = x).$$

- **Continuous:**

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

Product of Independent Random Variables

If X and Y are independent random variables:

$$E[XY] = E[X] \times E[Y].$$

Non-negative Random Variables

If $X \geq 0$, then:

$$E[X] \geq 0.$$

Jensen's Inequality

For a convex function g:

$$E[g(X)] \geq g(E[X]).$$

For a concave function, the inequality is reversed.
Example 1:

Let X be a random variable with $E[X] = 5$. Find $E[3X + 2]$.
Solution:

$$E[3X + 2] = 3E[X] + 2 = 3 \times 5 + 2 = 17.$$

7.2 Variance

Variance of a random variable is given by:

$$Var(X) = E((X - \mu)^2) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$$

Properties of Variance

Variance of a Constant

$$\text{Var}(c) = 0,$$

where c is a constant.

Scaling Property

For a constant a:

$$\text{Var}(aX) = a^2 \, \text{Var}(X).$$

Variance of a Linear Combination

For random variable X and constants a and b:

$$\text{Var}(aX + b) = a^2 \, \text{Var}(X).$$

Note: Adding a constant b does not affect the variance.

Variance of the Sum of Random Variables

For any two random variables X and Y:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\, \text{Cov}(X, Y),$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y.

Variance of Independent Random Variables

If X and Y are independent:

$$\mathrm{Cov}(X,Y)=0,$$

so:

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

Generalization to n Independent Random Variables

If $X_1, X_2, \ldots, X_n$ are independent random variables:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n}\mathrm{Var}(X_i).$$

Variance of the Difference of Random Variables

$$\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)-2\,\mathrm{Cov}(X,Y).$$

For independent X and Y:

$$\mathrm{Var}(X-Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

7.3 Standard Deviation

Standard deviation is the square root of variance:

$$\sigma=\sqrt{Var(X)}$$

Note

1. Expected value $\mu = E(X)$ is a measure of central tendency.
2. Standard deviation σ is a measure of spread.

7.4 Covariance and Correlation

Covariance

$$\mathrm{Cov}(X,Y)=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]E[Y].$$

Properties of Covariance

1. **Symmetry:**

$$\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X).$$

2. **Covariance with Itself:**

$$\mathrm{Cov}(X,X)=\mathrm{Var}(X).$$

3. **Linearity:**

$$\mathrm{Cov}(aX+b,Y)=a\,\mathrm{Cov}(X,Y).$$

$$\mathrm{Cov}(X,aY+b)=a\,\mathrm{Cov}(X,Y).$$

4. **Addition:**

$$\mathrm{Cov}(X_1+X_2,Y)=\mathrm{Cov}(X_1,Y)+\mathrm{Cov}(X_2,Y).$$

Correlation Coefficient

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y.

Properties:

- $-1 \leq \rho_{XY} \leq 1$.
- $\rho_{XY} = 1$ if $Y = aX + b$ with $a > 0$.
- $\rho_{XY} = -1$ if $Y = aX + b$ with $a < 0$.
- $\rho_{XY} = 0$ indicates no linear relationship (but not necessarily independence).

Example 2:

Let X and Y be independent random variables with $\text{Var}(X) = 4$ and $\text{Var}(Y) = 9$. Find $\text{Var}(2X - 3Y + 5)$.

Solution:

1. **Calculate Variance:**

$$\text{Var}(2X - 3Y + 5) = 2^2\,\text{Var}(X) + (-3)^2\,\text{Var}(Y) = 4 \times 4 + 9 \times 9 = 16 + 81 = 97.$$

Chebyshev's Inequality

For any random variable X with finite mean μ and variance σ^2, and for any $k > 0$:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Interpretation: The probability that X deviates from its mean by k standard deviations or more is at most $\frac{1}{k^2}$.

8 Discrete Distributions

8.1 Discrete Uniform Distribution

A discrete random variable defined for values of x from 1 to n is said to have a uniform distribution if its probability mass function is given by

$$f(x) = \frac{1}{n}, \quad \text{for } x = 1, 2, 3, \ldots, n$$
$$f(x) = 0, \quad \text{otherwise}$$

The cumulative distribution function $F(x)$ of the discrete uniform random variable x is given by:

$$F(x) = \begin{cases} \frac{x}{n}, & \text{for } 1 \leq x \leq n \\ 1, & \text{for } x > n \\ 0, & \text{for } x < 1 \end{cases}$$

The mean of $X = \mu$ is

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

The variance of $X = \sigma^2$ is

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

8.2 Binomial Distribution

An experiment which is made of n independent trials, each of which resulting in either 'success' with probability p or 'failure' with probability $q = 1 - p$, then the probability distribution for the random variable X when represents the number of success is called a binomial distribution. The probability mass function,

$$p(x) = b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

Properties of Binomial Distribution

1. $E(X) = np$ (mean)
2. $V(X) = E(X^2) - (E(X))^2 = npq$ (variance) (mean > variance)
3. $SD(X) = \sqrt{npq}$
4. Mode of a binomial distribution lies between $(n+1)p - 1 \leq x \leq (n+1)p$
5. If $X_1 \sim b(n_1, p)$ and $X_2 \sim b(n_2, p)$ and if X_1 and X_2 are independent, then $X_1 + X_2 \sim b(n_1 + n_2, p)$, where (n, p) is the pmf of binomial distribution.

8.3 Poisson Distribution

A random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$, if it assumes only non-negative values and its probability mass function is given by

$$p(x) = p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \quad \lambda > 0$$
$$p(x) = 0, \quad \text{otherwise}$$

In a binomial distribution if n is large compared to p, then np approaches a fixed constant λ. Such a distribution is called Poisson distribution (limiting case of binomial distribution).

Properties of Poisson Distribution

1. $E(X) = \sum_{x=0}^{\infty} x \cdot p(x) = \lambda$
2. $V(X) = E(X^2) - (E(X))^2 = \lambda$
3. $SD(X) = \sqrt{\lambda}$ (Mean = Variance)
4. Mode of a Poisson distribution lies between $\lambda - 1$ and λ
5. If $X_1 \sim P(\lambda_1)$ and $X_2 \sim P(\lambda_2)$, and X_1, X_2 are independent, then $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$

9 Continuous Distributions

9.1 Uniform/Rectangular Distribution

A continuous random variable x defined on $[a, b]$ is said to have a uniform distribution if its probability density function is given by

$$f(x) = \frac{1}{b-a}, \quad \text{for } x \in [a, b]$$
$$f(x) = 0, \quad \text{otherwise}$$

The cumulative distribution function of the continuous uniform random variable X is given by:

$$F(x) = \begin{cases} 0, & \text{for } x < a \\ \frac{x-a}{b-a}, & \text{for } a \leq x \leq b \\ 1, & \text{for } x > b \end{cases}$$

- Mean of $X = \mu = \frac{a+b}{2}$
- Variance of $X = \sigma^2 = \frac{(b-a)^2}{12}$

9.2 Normal Distribution

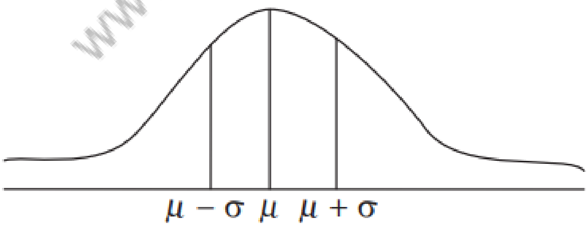
A continuous random variable X is said to follow a **normal distribution** with mean μ and variance σ^2 if its probability density function (pdf) is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

This is denoted as:

$$X \sim N(\mu, \sigma^2).$$

The graphical representation of normal distribution is as follows:



Note

- **Symmetry:** The normal distribution is symmetric about the mean μ.
- **Mean, Median, Mode:** All are equal and located at $x = \mu$.
- **Inflection Points:** The curve changes concavity at $x = \mu \pm \sigma$.
- **Standard Deviation:** Measures the spread of the distribution; larger σ means wider spread.
- **Total Area:** The total area under the curve is 1.

The 68-95-99.7 Rule

Approximately:

- 68% of the data falls within one standard deviation ($\mu \pm \sigma$).
- 95% within two standard deviations ($\mu \pm 2\sigma$).
- 99.7% within three standard deviations ($\mu \pm 3\sigma$).

9.3 Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. Its pdf is:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty.$$

We denote:

$$Z \sim N(0, 1).$$

Standardization

Any normal random variable $X \sim N(\mu, \sigma^2)$ can be transformed into a standard normal variable Z using:

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

This process is called **standardization** and allows us to use standard normal distribution tables to compute probabilities.

Example 1

Suppose $X \sim N(50, 16)$. Find $P(X > 58)$.

Solution:

1. **Standardize X:**

$$Z = \frac{X - \mu}{\sigma} = \frac{58 - 50}{4} = 2.$$

2. **Find $P(Z > 2)$:**

$$P(Z > 2) = 1 - P(Z \le 2) = 1 - 0.9772 = 0.0228.$$

(Using standard normal distribution table)

Example 2

Let $X_1, X_2, \ldots, X_{36}$ be independent and identically distributed random variables with mean $\mu = 10$ and variance $\sigma^2 = 9$. Find the approximate probability that the sample mean $\bar{X}$ exceeds 11.

Solution:

1. **Compute the standard deviation of $\bar{X}$:**

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{6} = 0.5.$$

2. **Standardize $\bar{X}$:**

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{11 - 10}{0.5} = 2.$$

3. **Find $P(\bar{X} > 11)$:**

$$P(Z > 2) = 0.0228.$$

Properties of Normal Distribution

1. The function is symmetrical about the value μ.
2. It has a maximum at $x = \mu$.
3. The area under the curve within the interval $(\mu \pm \sigma)$ is 68%.

$$\text{That is, } P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$$

4. A fairly large number of samples taken from a 'Normal' population will have average, median, and mode nearly the same, and within the limits of average $\pm 2 \times SD$, there will be 95% of the values.
5. $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \mu$.
6. $V(X) = \sigma^2$, $SD(X) = \sigma$.
7. For a normal distribution,

$$\text{Mean} = \text{Median} = \text{Mode}$$

8. All odd order moments about mean vanish for a normal distribution.

$$\mu_{2n+1} = 0 \quad \text{for} \quad n = 0, 1, 2, \ldots$$

9. If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, X_1, X_2 independent, then,

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Also,

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

10. If $\mu = 0$ and $\sigma^2 = 1$, we call it a standard normal distribution. The standardization can be obtained by the transformation,

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

Also,

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

9.4 Exponential Distribution

A continuous random variable X is said to have an exponential distribution if its probability density function $f(x)$ is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

Here λ is the parameter of the exponential distribution and $\lambda > 0$.
The cumulative distribution function $F(x)$ of an exponential distribution with λ as parameter is given by:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

The mean and variance for an exponential distribution are given by:

$$\text{Mean} = \mu = \frac{1}{\lambda}$$

$$\text{Variance} = \sigma^2 = \frac{1}{\lambda^2}$$

10 Joint Distribution of Random Variables

Let X and Y be two discrete random variables on the same sample space S with the range space of X as $R_x = \{x_1, x_2, \ldots, x_m\}$ and the range space of Y as $R_y = \{y_1, y_2, \ldots, y_n\}$ and $P_X(x)$ and $P_Y(y)$ as the probability mass functions of x and y. Then the joint probability mass function $P_{xy}(x, y)$ of the two dimensional random variable (x, y) on the range space $R_x \times R_y$ is defined as:

$$P_{XY}(x_i, y_j) = \begin{cases} P(X = x_i, Y = y_j), & \text{for } (x_i, y_j) \in R_x \times R_y \\ 0, & \text{otherwise} \end{cases}$$

This joint probability mass function can be represented in the form of a table as follows:

X	Y	y_1	y_2	$\cdots$	y_n	$\sum_{j=1}^{n} P_{xy}(x_i, y_j)$
x_1		$P_{xy}(x_1, y_1)$	$P_{xy}(x_1, y_2)$	$\cdots$	$P_{xy}(x_1, y_n)$	$P_X(x_1)$
x_2		$P_{xy}(x_2, y_1)$	$P_{xy}(x_2, y_2)$	$\cdots$	$P_{xy}(x_2, y_n)$	$P_X(x_2)$
x_3		$P_{xy}(x_3, y_1)$	$P_{xy}(x_3, y_2)$	$\cdots$	$P_{xy}(x_3, y_n)$	$P_X(x_3)$
$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m		$P_{xy}(x_m, y_1)$	$P_{xy}(x_m, y_2)$	$\cdots$	$P_{xy}(x_m, y_n)$	$P_X(x_m)$
$\sum_{i=1}^{m} P_{xy}(x_i, y_j)$		$P_Y(y_1)$	$P_Y(y_2)$	$\cdots$	$P_Y(y_n)$	

From the above table, it can be easily observed that the marginal probability mass functions of X and Y, namely $P_X(x)$ and $P_Y(y)$ respectively, can be obtained from the joint probability mass function $P_{xy}(x, y)$ as:

$$P_X(x_i) = \sum_{j=1}^{n} P_{xy}(x_i, y_j), \quad \text{for } i = 1, 2, \ldots, m$$

And

$$P_Y(y_j) = \sum_{i=1}^{m} P_{xy}(x_i, y_j), \quad \text{for } j = 1, 2, \ldots, n$$

- $P_{xy}(x_i, y_j) \geq 0 \, \forall \, i, j$
- $\sum_{i=1}^{m} \sum_{j=1}^{n} P_{xy}(x_i, y_j) = 1$
- The cumulative joint distribution function of the two dimensional random variable (X, Y) is given by $F_{xy}(x, y) = P(X \leq x, Y \leq y)$.

10.1 Joint Probability Density Function

Let X and Y be two continuous random variables on the same sample space S with $f_X(x)$ and $f_Y(y)$ as the probability density functions respectively. Then a function $f_{XY}(x, y)$ is called the joint probability density function of the two dimensional random variable (X, Y) if the probability that the point (x, y) will lie in the infinitesimal rectangular region of area $dxdy$ is $f_{XY}(x, y)\, dx\, dy$.

That is,

$$P\left(x - \frac{1}{2}dx \leq X \leq x + \frac{1}{2}dx,\; y - \frac{1}{2}dy \leq Y \leq y + \frac{1}{2}dy\right) = f_{XY}(x, y)\, dx\, dy$$

- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx\, dy = 1$
- The marginal probability density functions $f_X(x)$ and $f_Y(y)$ of the two continuous random variables X and Y are given by:
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$
- The cumulative joint distribution function $F_{XY}(x, y)$ of the two-dimensional random variable (X, Y) (where X and Y are any two continuous random variables defined on the same sample space) is given by:
$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y)\, dx\, dy$$

10.2 Conditional Probability Functions of Random Variables

Let X and Y be two discrete (continuous) random variables defined on the same sample space with joint probability mass (density) function $f_{XY}(x, y)$, then:

1. The conditional probability mass (density) function $f_{X|Y}(x|y)$ of X, given $Y = y$, is defined as:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}, \quad \text{where } f_Y(y) \neq 0$$

2. The conditional probability mass (density) function $f_{Y|X}(y|x)$ of Y, given $X = x$, is defined as:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad \text{where } f_X(x) \neq 0$$

10.3 Independent Random Variables

Two discrete (continuous) random variables X and Y defined on the same sample space with joint probability mass (density) function $P_{XY}(x, y)$ are said to be independent if and only if:

$$P_{XY}(x, y) = P_X(x)P_Y(y)$$

Where $P_X(x)$ and $P_Y(y)$ are the marginal probability mass (density) functions of the random variables X and Y respectively.

Note

If the random variables X and Y are independent, then

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b)P(c \leq Y \leq d)$$

11 Mean, Median, Mode and Standard Deviation

11.1 Mean

Mean of a data generally refers to the arithmetic mean of the data. However, there are two more types of mean which are geometric mean and harmonic mean. The different types of mean for a set of values and grouped data are given in the following table.

Type of Mean	**Set of Values**	**Grouped Data**
Arithmetic Mean	$\frac{\sum x_i}{n}$	$\frac{\sum f_i x_i}{\sum f_i}$
Geometric Mean	$(x_1 x_2 x_3 \ldots x_n)^{\frac{1}{n}}$	$\text{Antilog}\left(\frac{\sum f_i \log x_i}{\sum f_i}\right)$
Harmonic Mean	$\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$	$\frac{\sum f_i}{\sum \frac{f_i}{x_i}}$

11.2 Median

For an ordered set of values, the median is the middle value. If the number of values is even, median is taken as the mean of the middle two values. For grouped data, median is given by:

$$\text{Median} = L + \left(\frac{N/2 - C}{f}\right) h$$

where,

$L =$ Lower boundary of the median class, $N = \sum f_i$, $C =$ Cumulative frequency up to the class before the median class

$h =$ Width of the median class, $f =$ Frequency of the median class

Note

- Median does not take into consideration all the items.
- The sum of absolute deviations taken about median is least.
- Median is the abscissa of the point of intersection of the cumulative frequency curves.
- Median is the best suited measure for open end classes

11.3 Mode

For a set of values, mode is the most frequently occurring value. For grouped data, mode is given by:

$$\text{Mode} = L + \left(\frac{f_1 - f_{i-1}}{(2f_1 - f_{i-1} - f_{i+1})}\right) h$$

where,

$L =$ Lower boundary of the modal class, $f_1 =$ Frequency of the modal class (Highest frequency)

$f_{i-1} =$ Frequency of the class before the modal class, $f_{i+1} =$ Frequency of the class after the modal class, $h =$ Width of the

Note: Relation between Mean, Median and Mode:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

11.4 Standard Deviation

Standard deviation is given by:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{For a Set of Values}$$

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i} \quad \text{For Grouped Data}$$

The standard deviation σ can alternatively be calculated using the formula:

$$\sigma = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

This is a useful formula for computational purposes.

Note

1. The square of standard deviation is termed as **variance**.
2. Standard deviation (SD) is the least mean square deviation.
3. If each item is increased by a fixed constant, the SD does not alter or **SD is independent of change of origin**.
4. Standard deviation depends on each and every data item.
5. For a discrete series in the form $a, a + d, a + 2d, \ldots$ (Arithmetic Progression, AP), the standard deviation is given by:

   $$\text{SD} = d\sqrt{\frac{n^2 - 1}{12}}$$

   where n is the number of terms in the series.

t-distribution and chi-distribution

12 t-distribution

The t-distribution, also known as Student's t-distribution, is a probability distribution that is symmetric and bell-shaped like the normal distribution. However, it has heavier tails, meaning it is more prone to producing values that fall far from the mean. This distribution is used primarily when the sample size is small, or the population standard deviation is unknown.

12.1 Key Characteristics:

1. Shape:
    - Bell-shaped and symmetric around the mean, similar to the normal distribution.
    - Heavier tails, which means there is more variability in extreme values compared to the normal distribution.
2. Degrees of Freedom (df):
    - The shape of the t-distribution depends on the degrees of freedom (df), which are related to the sample size.
    - Lower df results in heavier tails (more spread out), while higher df makes the t-distribution closer to the normal distribution.
    - Degrees of freedom are typically equal to the sample size minus one (df = n - 1) for a one-sample t-test.
3. Used in:
    - Hypothesis testing (like t-tests) when:
      - The sample size is small ($n < 30$).
      - The population standard deviation is unknown.
    - Estimating the mean of a normally distributed population when the sample size is small.
    - Constructing confidence intervals for small samples.

12.1.1 Comparing the t-distribution and Normal Distribution:

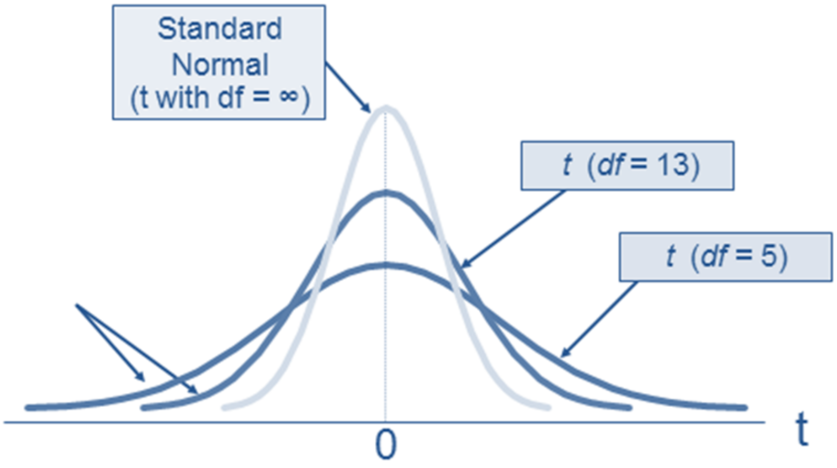
- Both are continuous probability distributions.
- The t-distribution is more spread out (has fatter tails) when the sample size is small, which accounts for the extra uncertainty in estimating the population standard deviation.
- As the sample size increases, the t-distribution gets closer to the normal distribution. When the sample size is large ($df > 30$), you can often use the normal distribution instead.

12.2 Formula for the t-distribution:

The probability density function of the t-distribution for a given degree of freedom df is given by:

$$f(x) = \frac{\Gamma\left(\frac{\mathrm{df}+1}{2}\right)}{\sqrt{\mathrm{df}\pi}\,\Gamma\left(\frac{\mathrm{df}}{2}\right)}\left(1+\frac{x^2}{\mathrm{df}}\right)^{-\frac{\mathrm{df}+1}{2}}$$

Where: - Γ is the gamma function (a generalization of factorial). - x is the value of the random variable.



12.3 Mean, Median, Mode and SD

1. Mean of t-distribution (df = 3): For a t-distribution:
   - The mean exists and is equal to **0** when the degrees of freedom (df) are greater than 1.
   - Since df = 3 > 1, the mean of the t-distribution is:

$$\text{Mean} = 0$$

2. Median of t-distribution (df = 3): - The t-distribution is symmetric about 0. Therefore, the median is the same as the mean.
   - The median of the t-distribution is:

$$\text{Median} = 0$$

3. Mode of t-distribution (df = 3): - The mode of a t-distribution is also located at 0, as the distribution is symmetric and centered at 0.
   - The mode of the t-distribution is:

$$\text{Mode} = 0$$

4. Variance of t-distribution (df = 3): - The variance of a t-distribution depends on the degrees of freedom (df). The variance is defined as:

$$\text{Variance} = \frac{df}{df - 2}$$

   - For df = 3:

$$\text{Variance} = \frac{3}{3-2} = \frac{3}{1} = 3$$

13 Chi-Square Distribution

The chi-square distribution is a statistical distribution commonly used in hypothesis testing, especially for categorical data and variance tests. It arises when you sum the squares of independent standard normal random variables. The chi-square distribution is primarily used in chi-square tests, which are applied to:
- Test goodness of fit (how well observed data fits a theoretical distribution).
- Test independence in contingency tables (e.g., association between two categorical variables).
- Test variance (whether the variance of a population equals a specified value).

13.1 Key Characteristics of the Chi-Square Distribution:

1. Non-negative Values: Since chi-square values are based on squared terms, the distribution is always non-negative.
2. Asymmetry: The chi-square distribution is right-skewed, especially for small degrees of freedom (df). As the degrees of freedom increase, the distribution becomes more symmetric and approaches a normal distribution.
3. Degrees of Freedom (df): The shape of the chi-square distribution depends on its degrees of freedom. The degrees of freedom typically represent the number of independent pieces of information available to estimate a parameter.

Chi-Square Distribution Formula:

If $X_1, X_2, \ldots, X_n$ are n independent normal variates with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then the standard normal variates are:

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{\mathrm{Var}(X_i)}} = \frac{X_i - \mu_i}{\sigma_i}, \quad i = 1, 2, \ldots, n$$

The sum of squares of standard normal variates is known as the chi-square variate with n degrees of freedom, i.e.,

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

If $Z_1, Z_2, ..., Z_k$ are independent, standard normal random variables (i.e., $Z_i \sim N(0,1)$), then the sum of their squares follows a chi-square distribution:

$$X^2 = Z_1^2 + Z_2^2 + ... + Z_k^2$$

The degrees of freedom (df) in this case are equal to k, the number of squared normal variables.

The probability density function is given by:

$$f(\chi^2) = \frac{1}{2^{k/2}\Gamma(k/2)} (\chi^2)^{k/2-1} e^{-\chi^2/2}, \quad 0 \leq \chi^2 < \infty$$

Applications of the Chi-Square Distribution:

1. Chi-Square Goodness-of-Fit Test:
   - Used to determine whether observed data fits a particular distribution.
   - Example: Testing whether a six-sided die is fair by comparing observed and expected frequencies of outcomes.

2. Chi-Square Test for Independence:
   - Used to determine whether two categorical variables are independent.
   - Example: Testing whether gender and voting preference are independent in a sample.

3. Chi-Square Test for Variance:
   - Used to test whether the variance of a population is equal to a specified value.
   - Example: Testing if the variance in test scores of students matches the expected variance.

13.2 Key Formulas for the Chi-Square Tests:

1. Chi-Square Goodness-of-Fit Test:

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

Where: - O_i = observed frequency, - E_i = expected frequency.
2. Chi-Square Test for Independence (Contingency Tables):

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where: - O_{ij} = observed frequency in the i, j-th cell, - E_{ij} = expected frequency in the i, j-th cell.
3. Chi-Square Test for Variance:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Where: - s^2 is the sample variance, - σ^2 is the population variance, - n is the sample size.

Probability Density Function (PDF) of Chi-Square Distribution
The chi-square distribution is a special case of the gamma distribution and its probability density function (PDF) is given by the following formula:

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}, \quad x > 0$$

Where:
- k is the degrees of freedom (df) of the distribution, - Γ is the gamma function, which generalizes the factorial function, - x is the random variable, which must be greater than 0.

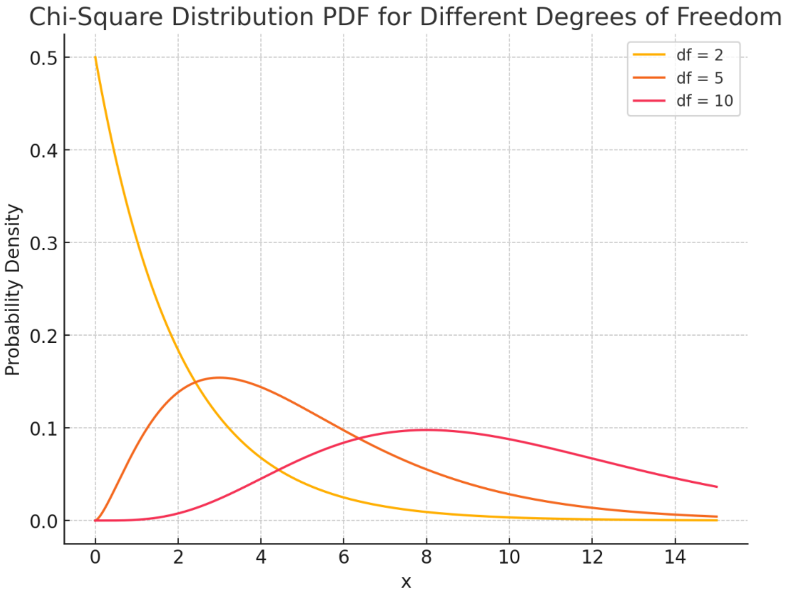
Key Properties of Chi-Square Distribution

- Mean: $\mu = k$
- Variance: $\sigma^2 = 2k$
- Skewness: The chi-square distribution is positively skewed (right-skewed), especially for smaller degrees of freedom. The skewness decreases as k increases.

Graph of Chi-Square Distribution (PDF)

The graph above shows the Probability Density Functions (PDF) of the chi-square distribution for different degrees of freedom ($df = 2, 5, 10$).

- For small degrees of freedom ($df = 2$), the distribution is highly skewed to the right.
- As the degrees of freedom increase ($df = 5$ and $df = 10$), the distribution becomes more symmetric and shifts to the right.



14 Hypothesis Testing

In probability theory, we set up mathematical models of processes and systems that are affected by 'chance'. In statistics, we check these models against the reality, to determine whether they are faithful and accurate enough for practical purposes. The process of checking models is called statistical inference. Methods of statistical inference are based on drawing samples (or sampling). One of the most important methods of statistical inference is 'Hypothesis Testing'.

Testing of Hypothesis

We have some information about a characteristic of the population which may or may not be true. This information is called statistical hypothesis or briefly hypothesis. We wish to know, whether this information can be accepted or to be rejected. We choose a random sample and obtain information about this characteristic. Based on this information, a process that decides whether the hypothesis to be accepted or rejected is called testing of hypothesis. i.e., In brief, the test of hypothesis or the test of significance is a procedure to determine whether observed samples differ significantly from expected results.

14.1 Key Terms:

1.Null Hypothesis (H_0):

- This is the default or initial assumption about the population.
- Example: $H_0 : \mu = 50$, meaning the population mean is equal to 50.

2.Alternative Hypothesis (H_a or H_1):

- Opposite of Null Hypothesis
- Example: $H_a : \mu \neq 50$, meaning the population mean is not equal to 50.

3.Test Statistic:

- A value calculated from the sample data that is used to decide whether to reject the null hypothesis. It measures how far the sample data diverge from the null hypothesis.
- Common test statistics include t-values (for t-tests), z-values (for z-tests), or chi-square values (for chi-square tests).

4.Significance Level (α):

- This is the threshold we set to determine how extreme the data must be to reject the null hypothesis. It is usually set to 0.05 (5%), meaning there's a 5% chance of rejecting the null hypothesis when it's actually true (Type I error).
- Example: If $\alpha = 0.05$, there's a 5% risk of incorrectly rejecting the null hypothesis.

5.p-value:

- The p-value is the probability of obtaining a test statistic at least as extreme as the one calculated from the data, assuming that the null hypothesis is true.
- If the p-value is less than or equal to the significance level ($p \leq \alpha$), you reject the null hypothesis.
- If the p-value is greater than the significance level ($p > \alpha$), you **fail to reject** the null hypothesis.

14.2 Steps in Hypothesis Testing:

1. State the hypotheses:
   - Null Hypothesis (H_0): This is the default claim you are testing.
   - Alternative Hypothesis (H_a): This is the claim you are trying to prove.
   - Example:
     - H_0: The average test score is 50.
     - H_a: The average test score is not 50.

2. Choose the significance level (α):
   - Typically, $\alpha = 0.05$ is used. It represents the probability of making a Type I error (rejecting H_0 when it is true).

3. Calculate the test statistic:
   - The test statistic depends on the type of data and the test you're conducting (t-test, z-test, etc.).
   - For a t-test, the formula is:

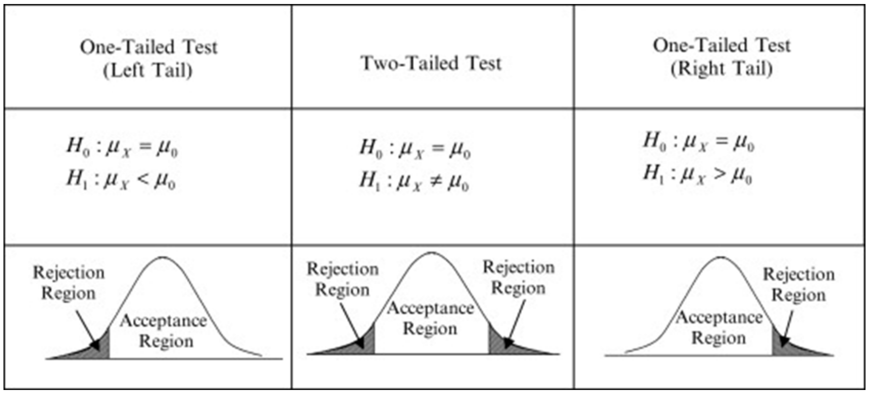
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

   Where: - $\bar{x}$ is the sample mean. - μ is the population mean under the null hypothesis. - s is the sample standard deviation. - n is the sample size.

4. Determine the p-value:
   - The p-value tells us how extreme the test statistic is under the assumption that H_0 is true.
   - The p-value is compared to α (significance level).

5. Make a decision:
   - If the p-value $\leq \alpha$: Reject the null hypothesis. There is evidence to support the alternative hypothesis.
   - If the p-value $> \alpha$: Fail to reject the null hypothesis. There is not enough evidence to support the alternative hypothesis.

6. Draw a conclusion: - Based on the results, you can conclude whether or not there is evidence to support the alternative hypothesis.



One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0: \mu_X = \mu_0$ $H_1: \mu_X < \mu_0$	$H_0: \mu_X = \mu_0$ $H_1: \mu_X \neq \mu_0$	$H_0: \mu_X = \mu_0$ $H_1: \mu_X > \mu_0$
Rejection Region Acceptance Region	Rejection Region Acceptance Region Rejection Region	Acceptance Region Rejection Region

Types of Hypothesis Tests:

1. One-Tailed Test: - Used when the alternative hypothesis is testing for a direction (greater than or less than). - Example: $H_a : \mu > 50$ (right-tailed) or $H_a : \mu < 50$ (left-tailed).
2. Two-Tailed Test: - Used when the alternative hypothesis is testing for any difference, not a specific direction. - Example: $H_a : \mu \neq 50$.

Errors in Hypothesis Testing:

- Type I Error (False Positive): - Occurs when you reject the null hypothesis when it is actually true. - The probability of making a Type I error is the significance level (α).
- Type II Error (False Negative): - Occurs when you fail to reject the null hypothesis when the alternative hypothesis is true.

14.3 Significance Level and Confidence Level in Terms of t-Value

In hypothesis testing and confidence interval estimation, significance level and confidence level are key concepts used to interpret results.

1. Significance Level (α)

The significance level (α) represents the probability of rejecting the null hypothesis when it is actually true. It is the threshold we set for determining how unlikely a result must be, under the null hypothesis, to reject that hypothesis. In other words, it's the probability of making a Type I error (false positive).

- A common significance level is $\alpha = 0.05$, meaning there is a 5% chance of incorrectly rejecting the null hypothesis when it is actually true.
- In a t-test, the significance level helps determine the critical value of the t-distribution, which is the cutoff point for deciding whether to reject the null hypothesis.

In Terms of t-Value:
- The t-critical value is the t-value that corresponds to the significance level in a t-distribution.
- For example, in a two-tailed test with $\alpha = 0.05$, the critical t-values are chosen so that 5% of the total area under the t-distribution curve is in the tails (2.5% in each tail). If the calculated t-value is more extreme than the critical value, we reject the null hypothesis.

Example:
For a two-tailed test with $\alpha = 0.05$ and $df = 9$:
- The critical values from the t-table are approximately ±2.262.
- If your calculated t-value is outside this range (less than -2.262 or greater than 2.262), you reject the null hypothesis.

2. Confidence Level

The confidence level is the percentage of confidence we have that the population parameter lies within the estimated range. It is complementary to the significance level:

$$\text{Confidence Level} = 1 - \alpha$$

For example: - If $\alpha = 0.05$, the confidence level is $1 - 0.05 = 0.95$

- This means we are 95% confident that the population parameter (such as the population mean) lies within the interval calculated from the sample data.

In Terms of t-Value:

- The confidence interval (CI) is calculated using the t-critical value based on the confidence level. For a 95% confidence interval, 95 of the data from the t-distribution will fall between the two t-critical values, leaving 5% in the tails (2.5% in each tail).

- To calculate the confidence interval for the population mean, we use:

$$CI = \bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

Where: - $\bar{x}$ is the sample mean, - $t_{\alpha/2}$ is the t-critical value corresponding to the confidence level, - s is the sample standard deviation, - n is the sample size.

Example:

For a sample of size 10, with $df = 9$, the t-critical value for a 95% confidence interval is approximately ±2.262. If the sample mean is 50 and the sample standard deviation is 4, the 95% confidence interval for the population mean is:

$$CI = 50 \pm 2.262 \times \frac{4}{\sqrt{10}} = 50 \pm 2.86$$

So the confidence interval is approximately $47.14 \leq \mu \leq 52.86$. This means we are 95% confident that the true population mean lies within this range.

t-test, z-test and chi-test

15 t-test

A t-value (or t-statistic) is a ratio that compares the difference between a sample statistic (like a sample mean) and a hypothesized population parameter (like a population mean) relative to the variability in the data. It is used in t-tests, a type of hypothesis test, when the population standard deviation is unknown and the sample size is small (usually $n < 30$).

The formula for calculating the t-value is:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where: - $\bar{x}$ is the sample mean, - μ is the hypothesized population mean, - s is the sample standard deviation, - n is the sample size.

The t-value tells you how many standard deviations the sample mean is from the hypothesized population mean.

Types of t-Tests:

1. One-Sample t-Test: Compares the sample mean to a known population mean.

   t-Statistic Calculation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

2. Two-Sample t-Test: Compares the means of two independent samples to see if they are significantly different.

   t-Statistic Calculation (assuming equal variances):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

   Where s_p^2 is the pooled variance:

3. Paired t-Test: Compares two related groups (like before and after measurements on the same group).

Properties of the t-Distribution:

- **Symmetry**: The t-distribution is symmetric and bell-shaped, just like the normal distribution.
- **Heavier Tails**: The t-distribution has heavier tails than the normal distribution, which means that it is more prone to producing values that fall far from the mean.
- **Degrees of Freedom (df):** The shape of the t-distribution depends on the degrees of freedom. As the degrees of freedom increase, the t-distribution approaches the normal distribution.
  - df for one-sample t-test: $n - 1$
  - df for two-sample t-test: $n_1 + n_2 - 2$
- **Mean:** The mean of the t-distribution is always 0.
- **Variance:** The variance is greater than 1 for small sample sizes but approaches 1 as the degrees of freedom increase.
- **Use for Small Samples:** The t-distribution is especially useful when the sample size is small and the population standard deviation is unknown.

How is the t-Value Used in Hypothesis Testing?

1. **Formulate Hypotheses: - Null Hypothesis (H_0)**: There is no effect or no difference. - **Alternative Hypothesis (H_a)**: There is an effect or a difference.

2. Calculate the t-Value: Use the formula for the t-statistic to calculate the t-value from the sample data.

3. **Determine the Critical Value:** Based on the significance level (usually $\alpha = 0.05$) and degrees of freedom, look up the critical value from the t-distribution table.

4. **Compare the t-Value and Critical Value:** - If the absolute value of the t-value is greater than the critical value, reject the null hypothesis.
   - If the absolute value of the t-value is less than the critical value, fail to reject the null hypothesis.

5. **Decision:** Based on the comparison, make a conclusion about whether there is enough evidence to support the alternative hypothesis.

Properties of the t-Value in t-Tests:

1. Sensitivity to Sample Size: The t-value is sensitive to the sample size. Larger samples result in a t-distribution that closely resembles the normal distribution.

2. Smaller Samples, More Variability: With smaller sample sizes, the t-distribution has more variability (heavier tails), and thus, a higher t-value is needed to reject the null hypothesis.

3. Depends on Sample Standard Deviation: The t-value incorporates the sample standard deviation, making it more appropriate for small samples where population variability is unknown.

4. Degrees of Freedom Influence: The degrees of freedom impact the shape of the t-distribution, with lower degrees of freedom resulting in heavier tails (larger critical values).

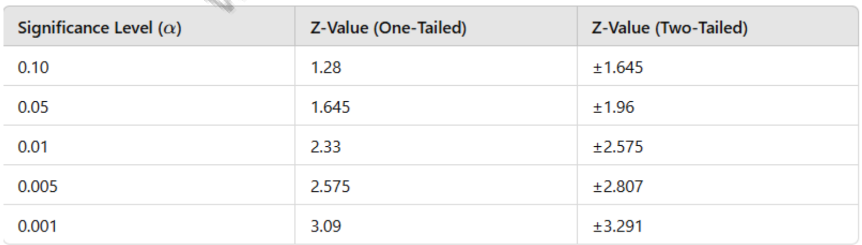
16 z-test

A Z-test is a statistical test used to determine whether there is a significant difference between sample and population means, or between two sample means, when the population variance is known or when the sample size is large (typically $n > 30$). The Z-test is based on the standard normal distribution (also called the Z-distribution), which has a mean of 0 and a standard deviation of 1.

Types of Z-Tests:

1. **One-sample Z-test:** Used to test if the sample mean is different from a known population mean.
   Formula:
   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
   Where: - $\bar{x}$ = sample mean, - μ = population mean, - σ = population standard deviation, - n = sample size.
2. **Two-sample Z-test** Used to compare the means of two independent samples.
   Formula:
   $$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
   Where: - $\bar{x}_1, \bar{x}_2$ = sample means, - σ_1^2, σ_2^2 = population variances, - n_1, n_2 = sample sizes.
3. **Z-test for proportions:** Used to compare proportions between two samples. **Formula:**
   $$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
   Where: - p_1, p_2 = sample proportions, - $\hat{p}$ = pooled proportion:
   $$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
   - n_1, n_2 = sample sizes.

Z-Test Critical Values Table



Significance Level (α)	Z-Value (One-Tailed)	Z-Value (Two-Tailed)
0.10	1.28	±1.645
0.05	1.645	±1.96
0.01	2.33	±2.575
0.005	2.575	±2.807
0.001	3.09	±3.291

- **One-Tailed Test:** Use this when you're testing for a directional hypothesis (e.g., whether a sample mean is greater than or less than a population mean). The critical value here corresponds to the tail on one side of the normal distribution.

  Example: If you are conducting a one-tailed test at $\alpha = 0.05$, the Z-critical value is 1.645.
- **Two-Tailed Test:** Use this when you're testing for a non-directional hypothesis (e.g., whether the sample mean is different from the population mean, either higher or lower). This splits the significance level equally between both tails of the normal distribution.

  Example: If you're conducting a two-tailed test at $\alpha = 0.05$, the Z-critical values are ± 1.96.

17 chi-test

The Chi-Square Test (or χ^2 test) is a statistical test used to determine if there is a significant association between observed and expected frequencies or whether two categorical variables are independent. It is widely used in hypothesis testing for categorical data, especially when working with contingency tables and goodness-of-fit tests.

Types of Chi-Square Tests

1. Chi-Square Goodness-of-Fit Test: - Used when you want to determine if an observed frequency distribution fits a particular theoretical distribution. - Example: Testing whether a die is fair by comparing observed frequencies of each outcome with the expected frequencies.
2. Chi-Square Test for Independence: - Used when you want to determine if two categorical variables are independent of each other. - Example: Testing whether gender is independent of voting preference in a population using a contingency table.

Chi-Square Test Formula
The chi-square test statistic (χ^2) is calculated as follows:

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

Where: - O_i = Observed frequency, - E_i = Expected frequency.

Key Assumptions:

1. Data must be in the form of frequencies (counts).
2. The observations must be independent of each other.
3. The sample size should be large enough (expected frequencies should ideally be 5 or more).

1. Chi-Square Goodness-of-Fit Test

This test compares the observed distribution of data to a theoretically expected distribution. The test checks how well the observed data "fits" the expected data.

Steps for Chi-Square Goodness-of-Fit Test:

1. Hypotheses:
   - Null Hypothesis (H_0): The observed data fits the expected distribution.
   - Alternative Hypothesis (H_a): The observed data does not fit the expected distribution.
2. Degrees of Freedom (df):
   - $df = k - 1$, where k is the number of categories.
3. Test Statistic:
   - Calculate the chi-square statistic using the formula:

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

4. Compare to Critical Value:
   - Determine the critical value from the chi-square distribution table using the degrees of freedom and the significance level α.
5. Decision:
   - If the calculated chi-square value is greater than the critical value, reject H_0. - Otherwise, fail to reject H_0.

2. Chi-Square Test for Independence

This test checks whether two categorical variables are independent of each other. It's often used in a contingency table, which shows the frequency distribution of variables.

Steps for Chi-Square Test for Independence:

1. Hypotheses: - Null Hypothesis (H_0): The two variables are independent.
   - Alternative Hypothesis (H_a): The two variables are dependent.

2. Expected Frequency:
   - Calculate the expected frequency for each cell in the contingency table:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

3. Degrees of Freedom (df):
   - $df = (r-1) \times (c-1)$, where r is the number of rows, and c is the number of columns.

4. Test Statistic:
   - Calculate the chi-square statistic using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

5. Compare to Critical Value:
   - Find the critical value from the chi-square table using the degrees of freedom and the significance level α.

6. Decision:
   - If χ^2 exceeds the critical value, reject H_0. - Otherwise, fail to reject H_0.

Important Considerations:

- Expected Frequencies: Each expected frequency should ideally be 5 or more for the chi-square test to be valid.
- Non-Negativity: Chi-square values are always non-negative because they are based on squared differences.
- Right-Skewed Distribution: The chi-square distribution is skewed to the right, especially for small degrees of freedom.

3. Chi-Square Test for Variance

The Chi-Square Test for Variance is a statistical method used to test if the variance (or standard deviation) of a population is equal to a specified value. This test is especially useful when we want to determine if a sample variance differs significantly from a known or hypothesized population variance.

The chi-square distribution is used because the test statistic for the variance follows a chi-square distribution if the population is normally distributed.

Formula for Chi-Square Test for Variance

The test statistic for the chi-square test for variance is:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Where: - χ^2 = Chi-square test statistic, - n = Sample size, - s^2 = Sample variance, - σ^2 = Hypothesized population variance.

Key Hypotheses for the Test

1. **Null Hypothesis (H_0):** The population variance is equal to a specific value.

$$H_0 : \sigma^2 = \sigma_0^2$$

2. Alternative Hypothesis (H_a): The population variance is not equal to the specified value (can be one-sided or two-sided).

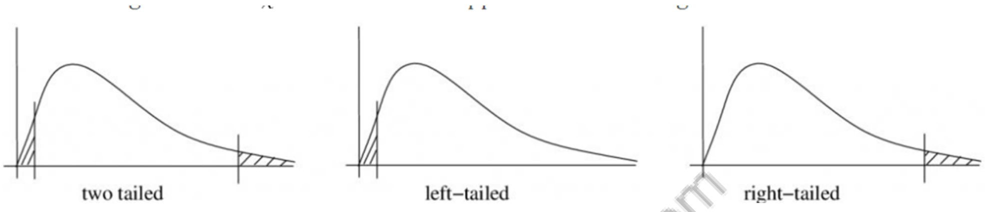
$$H_a : \sigma^2 \neq \sigma_0^2 \quad \text{(for two-tailed test)}$$

$$H_a : \sigma^2 > \sigma_0^2 \quad \text{(for right-tailed test)}$$

$$H_a : \sigma^2 < \sigma_0^2 \quad \text{(for left-tailed test)}$$

The possible hypothesis pairs are, for variance :

Two-tailed Test	Right-tailed Test	Left-tailed Test
$H_0 : \sigma^2 = k$	$H_0 : \sigma^2 \leq k$	$H_0 : \sigma^2 \geq k$
$H_1 : \sigma^2 \neq k$	$H_1 : \sigma^2 > k$	$H_1 : \sigma^2 < k$



Steps for Conducting the Chi-Square Test for Variance

1. State the Hypotheses:
   - Define the null and alternative hypotheses based on the problem.
2. Set the Significance Level (α):
   - Common values for α are 0.05 or 0.01.
3. Calculate the Test Statistic:
   - Use the formula:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

4. Degrees of Freedom (df):
   - $df = n - 1$, where n is the sample size.
5. Find the Critical Value:
   - Use the chi-square distribution table to find the critical value based on the significance level α and degrees of freedom df.
6. Make a Decision:
   - Compare the calculated chi-square value with the critical value(s):
   - For a two-tailed test, if the calculated χ^2 value falls outside the range of critical values, reject H_0.
   - For a one-tailed test, if the calculated χ^2 value is greater than the upper critical value (or less than the lower critical value for a left-tailed test), reject H_0.
7. Conclusion:
   - Based on the decision, either reject or fail to reject the null hypothesis.

18 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental statistical concept that states that the sampling distribution of the sample mean (or sum) approaches a normal distribution as the sample size becomes large, regardless of the shape of the population distribution. This theorem is crucial because it allows us to make inferences about population parameters even when the population distribution is not normal.

18.1 Key Points of the Central Limit Theorem:

1.Applies to Sample Means:

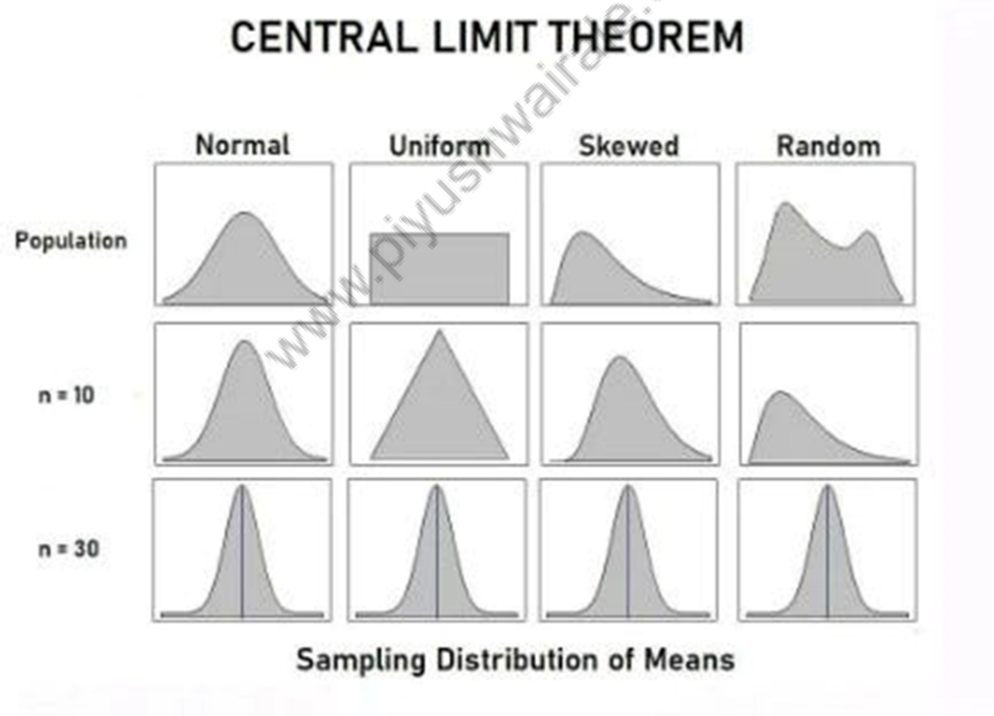
- The CLT applies to the distribution of the sample mean, $\bar{x}$, from a population. As the sample size n increases, the distribution of $\bar{x}$ approaches a normal distribution, even if the population itself is not normally distributed.

Sample Size:

- The CLT holds when the sample size is sufficiently large. In practice, a sample size of 30 or more is generally considered large enough for the CLT to apply, though this can vary depending on the shape of the population distribution.

3.Normal Distribution:

- Regardless of the population distribution (skewed, uniform, etc.), the distribution of the sample means becomes approximately normal as the sample size increases.



4.Mean and Standard Deviation:

- The mean of the sampling distribution of the sample mean is equal to the population mean (μ). - The standard deviation of the sampling distribution (also called the standard error) is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where: - σ = population standard deviation, - n = sample size.

Statement of the Central Limit Theorem:

If you take a large enough sample size n from any population with mean μ and standard deviation σ, the distribution of the sample means will be approximately normal, with: - Mean $= \mu$, - Standard error $= \frac{\sigma}{\sqrt{n}}$.

Mathematically, for large n, the sample mean $\bar{x}$ follows a normal distribution:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

References

- Sheldon Ross, *A First Course in Probability*, 9th Edition.
- S. C. Gupta and V. K. Kapoor, *Fundamentals of Mathematical Statistics.*
- William Feller, *An Introduction to Probability Theory and Its Applications.*