

Chapter #2

“Mathematics for Data Science”

10 Hours | 12 Marks

Compiled By
Er. Sushil Dyopala

Topics

1. Introduction to linear algebra for data science
2. Vectors, matrices and matrix factorization
3. Gradient descent for optimization
4. Introduction to probability and random variable
5. Probability distributions: Normal, Bernoulli, Binomial, Poisson
6. Descriptive and inferential statistics
7. Central limit theorem and sample distribution concepts
8. Normal approximation; hypothesis testing procedures: Tests about the mean of a normal population
9. The t-test, Z-tests for differences between two populations means, the two sample t-test, confidence interval for mean of normal population
10. ANOVA

1.0 What is Linear Algebra?

- Linear algebra is the branch of mathematics that deals with **matrices**, **vectors**, **vector spaces**, and **linear transformations**.
- Linear algebra forms the foundation of many mathematical, computational, and engineering fields, including data science, machine learning, computer graphics, physics, and economics.
- It is a tool to work with multidimensional data.
- Key Concepts Used
 - Vectors and vector spaces
 - Systems of linear equations
 - Matrices
 - Eigenvalues and Eigenvectors

1.1 Applications: How is Linear Algebra used in DS?

- **Data Representation**

- Data sets are often represented as matrices, wherein every row corresponds to an observation and every column represents a feature.
- This matrix illustration permits efficient manipulation and data analysis.

- **Dimensionality Reduction**

- Techniques Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) methods rely on principles from linear algebra to decrease the complexity of data while retaining critical information.

- **Optimization**

- Linear algebra is important for optimization algorithms utilized in machine learning, including gradient descent, based on calculating gradients.

- **Feature Engineering**

- Linear transformations and matrix operations create new features from existing data

1.1 Applications: How is Linear Algebra used in DS?

- **Machine Learning Algorithms**

- Algorithms like support vector machines, linear regression, and logistic regression utilize linear algebra operations to build models.

- **Similarity Measures**

- Embeddings are stored as vectors and are used in Natural Language Processing (eg. chatbots)

- **Eigenvalues and Eigenvectors**

- These concepts assist in identifying dominant patterns and directions of variability in data, useful in clustering, and feature extraction.

- **Image and Signal Processing**

- Linear algebra strategies are vital in image processing tasks like filtering, compression, and edge detection.
- Fourier transforms, and convolution operation contain linear algebra operations.

2.1 Vectors: Introduction

- A vector is an ordered list of numbers, which represents a point, and a direction in a multidimensional space.
- In data science, vectors often represent features or observations in datasets.
- The number of elements in a vector determines its dimensionality. If a vector has 3-elements, it will be a 3-dimensional vector.
- Vectors are typically written as:

- Column vector
- Row vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}.$$

m-dimensional column vector
(m x 1)

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$$

m-dimensional row vector
(1 x m)

2.1 Vectors: Magnitude & Direction

- **Magnitude or Length of Vector**

- Magnitude of a vector represents its size or length and is denoted as $||v||$ or $|v|$.
- Magnitude of a vector $\mathbf{v} = [v_1, v_2, \dots, v_n]$ is calculated using the Pythagorean theorem:

$$||\mathbf{v}|| = \sqrt{(v_1^2 + v_2^2 + \dots + v_n^2)}$$

- Magnitude of vectors is also called **norm**. Above formula represents 2-norm i.e. Euclidean Distance.

- **Direction of Vector**

- The direction of an n-dimensional vector describes the orientation of the vector in space.
- It is often expressed as a unit vector, which has the same direction but a magnitude of 1.

2.1 Vectors: Unit Vector

- A unit vector is a vector with a magnitude of 1.
- It is often used to represent the direction of a vector without regard to its magnitude.
- They serve as a normalized form of a given vector and are particularly useful in calculations involving directions and projections.
- Given a vector $\mathbf{v} = [v_1, v_2, \dots, v_n]$, the unit vector in the direction of \mathbf{v} is calculated by dividing each component of \mathbf{v} by its magnitude:

$$\mathbf{v}_{\text{unit}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

where, $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}$

- Properties:
 - Magnitude: $\|\mathbf{v}\| = 1$
 - The unit vector retains the direction of the original vector \mathbf{v} .
 - The unit vector has the same number of components as the original vector.

1. 2D Vector:

Let $\mathbf{v} = [3, 4]$. The magnitude of \mathbf{v} is:

$$\|\mathbf{v}\| = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = 5$$

The unit vector is:

$$\mathbf{v}_{\text{unit}} = \frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{[3, 4]}{5} = \left[\frac{3}{5}, \frac{4}{5}\right] = [0.6, 0.8]$$

2.1 Vectors: Orthogonal & Orthonormal Vectors

- **Orthogonal Vectors**

- Vectors are said to be orthogonal if they are perpendicular to each other.
- Their dot products will be 0 i.e. $\mathbf{u} \cdot \mathbf{v} = 0$
- They can have any magnitude (they are not required to have unit length).
- Eg: $\begin{bmatrix} 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 2 \end{bmatrix}$ are orthogonal vectors.

- **Orthonormal Vectors**

- Vectors are said to be orthonormal if they are orthogonal to each other and each vector has a magnitude of 1 (unit length).
- Mathematically, $\mathbf{u} \cdot \mathbf{v} = 0$ and $\|\mathbf{u}\| = 1, \|\mathbf{v}\| = 1$
- Eg: $\begin{bmatrix} 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \end{bmatrix}$ are orthonormal vectors.

2.1 Vector Operations

- **Addition:** Vectors can be added component-wise. If you have two vectors \mathbf{u} and \mathbf{v} , their sum \mathbf{w} is given by :

$$\mathbf{w} = \mathbf{u} + \mathbf{v}$$

- **Scalar Multiplication:** A vector can be scaled (multiplied) by a scalar (a single number). If \mathbf{v} is a vector and c is a scalar, then $c\mathbf{v}$ is a scaled version of \mathbf{v} .
- **Dot Product:** The dot product of two vectors \mathbf{u} and \mathbf{v} is a scalar quantity given by

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\theta)$$

where, θ is the angle between the two vectors.

The dot product measures the similarity or correlation between vectors.

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{bmatrix}$$

$$c \cdot \mathbf{v} = \begin{bmatrix} c \cdot v_1 \\ c \cdot v_2 \\ \vdots \\ c \cdot v_n \end{bmatrix}$$

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^n u_i v_i$$

$$\begin{pmatrix} 3 \\ 7 \end{pmatrix} \cdot \begin{pmatrix} 9 \\ 5 \end{pmatrix} = 3 * 9 + 7 * 5 = 62$$

2.2 Matrices: Introduction

- A matrix is a rectangular array of numbers, symbols, or expressions arranged in rows and columns.
- Matrices are denoted with uppercase letters (e.g. ***A***) and elements within a matrix are often denoted as ***a_{ij}***

General Representation:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Here ***A*** is the matrix of dimension ***m x n***

2.2 Matrices: Types

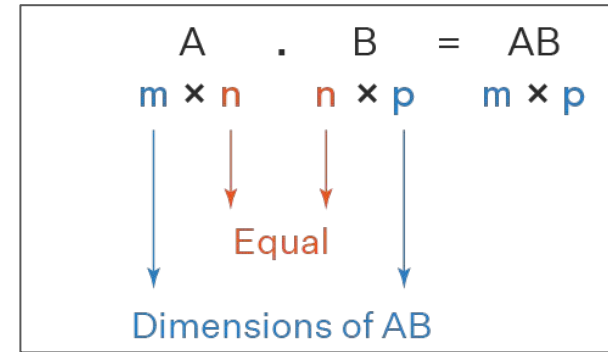
- Row Matrix: A matrix with a single row. Example: [1,2,3]
- Column Matrix: A matrix with a single column.
- Square Matrix: A matrix where the number of rows equals the number of columns ($m=n$).
- Diagonal Matrix: A square matrix where all elements outside the main diagonal are zero.
- Identity Matrix (I): A diagonal matrix with ones on the diagonal.
- Zero or Null Matrix: A matrix where all elements are zero.
- Upper Triangular Matrix: A square matrix where the elements below the principal diagonal are zero.
- Lower Triangular Matrix: A square matrix where the elements above the principal diagonal are zero.
- Singular Matrix: A matrix is said to be a singular matrix if its determinant $|A| = 0$
- **Orthogonal Matrix**: A square matrix whose rows and columns are **orthonormal vectors**. Each row and column is a unit vector. Rows are perpendicular (orthogonal) to each other, as are columns. Transpose of orthogonal matrix equals to its inverse. Orthogonal matrices preserve the lengths of vectors and the angles between them during transformations. They represent rotations or reflections in space.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xrightarrow{\text{Transpose } A^T} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \xleftarrow{\text{Inverse } A^{-1}}$$

Orthogonal Matrix

2.2 Matrix Operations

- Addition
$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{bmatrix}$$
- Subtraction
$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} - \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 - a_2 & b_1 - b_2 \\ c_1 - c_2 & d_1 - d_2 \end{bmatrix}$$
- Multiplication
$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \times \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{bmatrix}$$
 - For multiplication compatibility, columns of the first matrix must be equal to the rows of the second matrix. [**@C ROW**]
 - Matrix multiplication is not commutative.
 - $$C_{ij} = \sum_k A_{ik} B_{kj}$$



Matrix Multiplication Rules

2.2 Matrix Operations

- Transpose of Matrix: $[A(a_{ij}) = A^T(a_{ji})]$

$$\text{If a matrix } A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$\text{Transpose } A^t = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}$$

- Determinant of Matrix

- The determinant of a square matrix is a scalar value that provides information about the matrix like matrix invertibility, transformations, and system solvability.
- The determinant is only defined for square matrices.
- It is denoted as $\det(A)$ or $|A|$.
- A matrix is said to be singular if its determinant is zero.
- The determinant represents the scaling factor of the linear transformation described by the matrix.

2.2 Matrix Operations

- Inverse of Matrix

$$A^{-1} = \frac{1}{|A|} \cdot \text{Adj } A$$

- Property of Inverse Matrix:

Diagram illustrating the property of the inverse matrix:

Matrix A (blue) and Inverse of Matrix A (red) are multiplied to yield the Identity Matrix (I).

$$A \cdot A^{-1} = I$$

2.3 Data Representation Using Vectors and Matrices

- Illustration will be performed in class.

2.4 Linear Transformation

- A linear transformation in linear algebra is a mathematical function (T) that maps vectors from one vector space to another while preserving two key properties:
 - **Additivity** (or linearity):
 - $T(u+v) = T(u) + T(v)$ For all vectors u and v in the domain.
 - **Scalar multiplication**:
 - $T(c \cdot u) = c \cdot T(u)$ for any scalar c and vector u .
- Examples: rotation, reflection, scaling

2.4 Linear Transformation Example: Rotation in 2D

- Rotation matrix [for θ angle of rotation]:

$$R = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

- First verifying additive property:

Let:

$$\mathbf{u} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The sum is:

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Rotation transformation matrix R for 90° is:

$$R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Step 1: Rotate $\mathbf{u} + \mathbf{v}$:

$$T(\mathbf{u} + \mathbf{v}) = R \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Step 2: Rotate \mathbf{u} and \mathbf{v} separately:

$$T(\mathbf{u}) = R \cdot \mathbf{u} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$T(\mathbf{v}) = R \cdot \mathbf{v} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

Adding these:

$$T(\mathbf{u}) + T(\mathbf{v}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Result:

$$T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$$

2.4 Linear Transformation Example: Rotation in 2D

- Verifying Scalar Multiplication Property

Let:

$$\mathbf{u} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = 2$$

Step 1: Scale \mathbf{u} by c and then apply T :

$$c \cdot \mathbf{u} = 2 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$T(c \cdot \mathbf{u}) = R \cdot \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

Step 2: Apply T to \mathbf{u} , then scale by c :

$$T(\mathbf{u}) = R \cdot \mathbf{u} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$c \cdot T(\mathbf{u}) = 2 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

Result:

$$T(c \cdot \mathbf{u}) = c \cdot T(\mathbf{u})$$

Since 2D Rotation satisfies both additive and scalar multiplication properties, it is a linear transformation.

2.5 Matrix Factorization

- Matrix factorization is a mathematical technique that is used to decompose a matrix into the product of two or more smaller matrices.
- This process simplifies the representation of the original matrix and is widely used in areas such as linear algebra, machine learning, and data science.
- Also known as Matrix Decomposition.
- Given a matrix ***M*** of size ***m*×*n***, matrix factorization aims to approximate it as:

$$\mathbf{M} \approx \mathbf{U} \cdot \mathbf{V}$$

- where,
 - *U* is a matrix of size *m*×*k*,
 - *V* is a matrix of size *k*×*n*,
 - *k* is a parameter (rank) that is usually much smaller than *m* and *n*

2.5 Matrix Factorization: Applications & Importance

- **Dimensionality Reduction:**
 - It reduces a high-dimensional matrix into lower dimensions while preserving key information.
- **Latent Features:**
 - It helps discover hidden (latent) features in data.
- **Data Imputation:**
 - Missing values in the matrix can be predicted based on the factorized matrices.
- **Efficiency:**
 - Smaller matrices are easier to work with computationally.
- **Recommendation Systems:**
 - Used for Collaborative filtering (e.g., Netflix, Spotify).
- **Image Compression:**
 - Reducing storage space for images.

2.5 Matrix Factorization Techniques

- **Singular Value Decomposition (SVD)**

- SVD decomposes a matrix M into three components:

$$M = U \cdot \Sigma \cdot V^T$$

- where,
 - U : Orthogonal matrix of size $m \times m$,
 - Σ : Diagonal matrix of singular values of size $m \times n$,
 - V : Orthogonal matrix of size $n \times n$.
- SVD is widely used in dimensionality reduction and image compression.

- **Non-Negative Matrix Factorization (NMF)**

- NMF decomposes a matrix M such that all elements of U and V are non-negative:

$$M = U \cdot V$$

- This is especially useful when the data has inherent non-negative constraints, such as pixel intensities or user preferences.

2.5 Matrix Factorization Techniques

- **LU Decomposition**

- Decomposes a square matrix M into:

$$M = L \cdot U$$

- where:
 - L : Lower triangular matrix,
 - U : Upper triangular matrix.
- This is often used to solve systems of linear equations.

- **Eigen Decomposition**

- Decomposes a square matrix M into:

$$M = Q \cdot \Lambda \cdot Q^{-1}$$

- where:
 - Q : Matrix of eigenvectors,
 - Λ : Diagonal matrix of eigenvalues.

2.6 Matrix Factorization Example: Recommendation System

- Suppose we have a data set which contains the items ratings given by various users.
- We have to recommend new item to the specific user based on his previous ratings. For this we need to decompose the rating matrix.
- Let's take $n \rightarrow$ number of users, $m \rightarrow$ number of items then our Rating Matrix will be of the order of $(n \times m)$.

	Item			
	W	X	Y	Z
User A		4.5	2.0	
User B	4.0		3.5	
User C		5.0		2.0
User D		3.5	4.0	1.0

Rating Matrix
 $(n \times m)$

2.6 Matrix Factorization Example: Recommendation System

- After applying Matrix Factorization we get two low ranked matrices, user matrix of shape (nxk) and item matrix of shape (kxm). [k is the dimension of feature vector chosen during matrix factorization]
- The values of user matrix and Item matrix are optimized through appropriate objective functions.
- Using the user matrix and item matrix we can recommend new items to the user's preference and also we can recommend items to the user based on the preference of other users having same preference.

		Item			
		W	X	Y	Z
User	A		4.5	2.0	
	B	4.0		3.5	
	C		5.0		2.0
	D		3.5	4.0	1.0

Rating Matrix

=

A	1.2	0.8
B	1.4	0.9
C	1.5	1.0
D	1.2	0.8

User Matrix

X

	W	X	Y	Z
	1.5	1.2	1.0	0.8
	1.7	0.6	1.1	0.4

Item Matrix

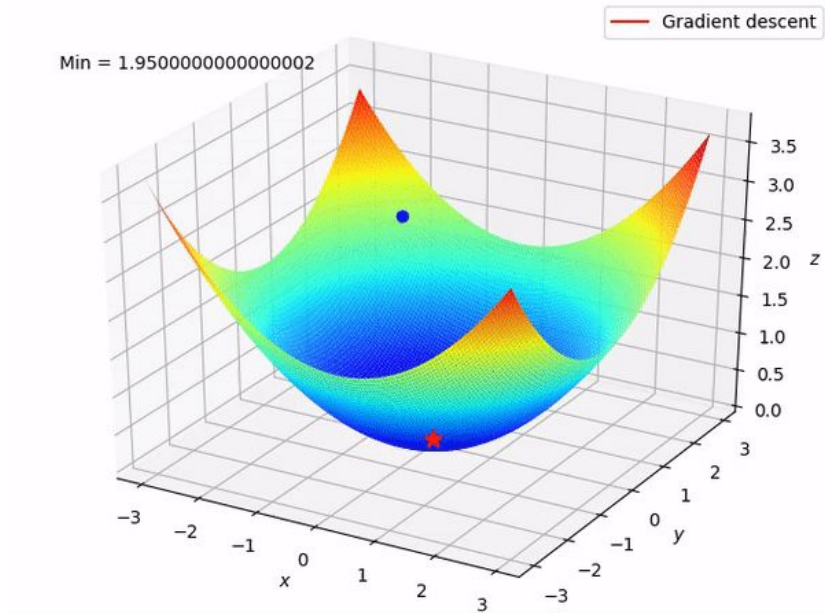
3.1 What is Optimization?

- Optimization is the process of finding the **best solution** either maximum or minimum **for a given objective function** by systematically adjusting parameters of a model within a defined range.
- It is one of the most important phenomena in Machine Learning.
- Objective function is the function to be optimized (minimized or maximized).
- In machine learning, objective function is to minimize loss or error or cost function. And optimization is used to minimize loss functions, which results in better models.
- **Gradient Descent** is the most commonly used Optimization Algorithm.

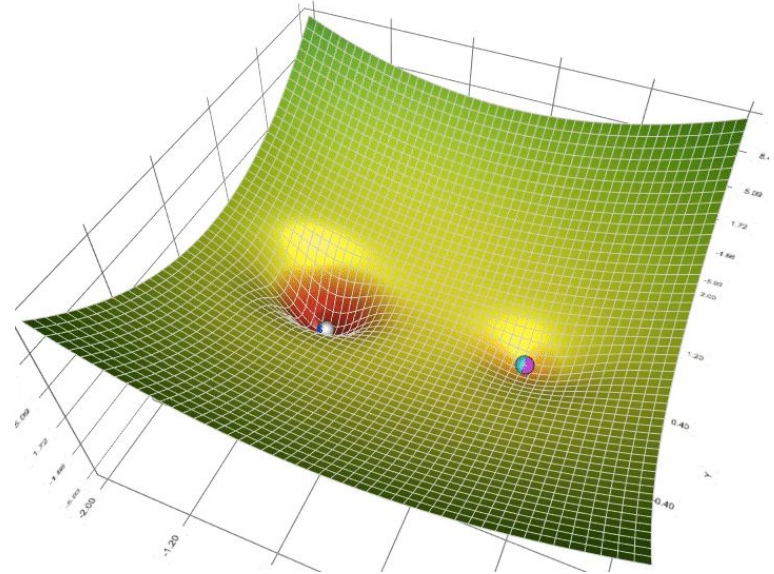
3.2 Gradient Descent Optimization Algorithm

- Gradient Descent is a **first-order iterative** optimization algorithm used to minimize functions by moving in the direction of the steepest descent (negative gradient).
- It is a minimization algorithm that minimizes a given function.
- Gradient descent algorithm works only for convex and differentiable functions.
- It finds out the **local minima** of a **differentiable function**.

3.2 Gradient Descent Optimization Algorithm



Gradient Descent in Convex Objective Function



Gradient Descent in Non-Convex Objective Function

3.2 Gradient Descent Optimization Algorithm

1. **Initialize Parameters:** Start with random values for the model's parameters (weights and biases) i.e. θ s
2. **Compute the Loss or Cost or Error:** Use the cost or loss or error function $J(\theta)$ to measure the error between the predicted output and the actual output.
3. **Calculate Gradients:** Compute the partial derivatives of the loss function with respect to the parameters to determine the gradient $\nabla_{\theta} J(\theta)$.
4. **Update Parameters:** Adjust the parameters using the formula:

$$\Theta = \theta - \alpha \cdot \nabla_{\theta} J(\theta)$$

where, θ : Parameters (e.g., weights)

α : Learning rate (step size)

$\nabla_{\theta} J(\theta)$: Gradient of the loss function

5. **Repeat:** Continue iterating from step 2 to 4 until convergence (the loss function stops decreasing significantly).

3.2 Gradient Descent Optimization Algorithm

- Algorithm Implementation and Visualization will be done in class.

4.1 Introduction to Probability and random variable

- Probability is a branch of mathematics that deals with the likelihood of occurrence of a given event.
- In data science, it is used to model uncertainty, make predictions, and infer patterns from data.
- Value of Probability ranges from 0 to 1.

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

4.1 Probability Terminologies

- **Experiment:** A procedure that yields one of many outcomes (e.g., rolling a die).
- **Sample Space (S):** The set of all possible outcomes of a random experiment. For example, in a coin toss, the sample space is {Heads, Tails}.
- **Event (E):** A subset of the sample space. An event occurs if the outcome of the experiment is in the subset.
 - Example: Rolling a six-sided die.
 - Sample Space (S): {1, 2, 3, 4, 5, 6}
 - Event (E): Rolling an even number, {2, 4, 6}

4.2 Random Variable

- A random variable is a function that assigns a numerical value to each outcome in the sample space of a random experiment.
- Example Experiment: Tossing two coins.
- Sample Space: S or $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$.
- Random Variable X : **Number of heads**
- Mapping:
 - $(H, H) \rightarrow X = 2$
 - $(H, T) \rightarrow X = 1$
 - $(T, H) \rightarrow X = 1$
 - $(T, T) \rightarrow X = 0$
- Thus, the possible values of X are $\{0, 1, 2\}$ with corresponding probabilities:
 - $P(X = 2) = 1/4$
 - $P(X = 1) = 2/4$
 - $P(X = 0) = 1/4$
- Mathematical Notation: If X is a random variable and x is a possible value, the probability is denoted as **$P(X=x)$**

4.2 Types of Random Variable

- **Discrete Random Variable**

- Takes on a finite or countable number of distinct values.
- Example: The number of heads in 10 coin tosses.
- Values: $\{0, 1, 2, \dots, 10\}$
- Probability Mass Function (PMF) is used to describe probability distribution of discrete random variable.

- **Continuous Random Variable**

- Takes on any value within a continuous range.
- Example: The time it takes for a website to load.
- Values: Any real number in the range $[0, \infty)$.
- Probability Density Function (PDF) is used to describe probability distribution of continuous random variable.

4.3 Probability Distributions

- Probability distributions describe how the probabilities are distributed over the values of the random variable.
- **Discrete Probability Distributions**
 - Applicable when the random variable takes on a finite or countable number of values.
 - Examples:
 - Bernoulli Distribution
 - Binomial Distribution
 - Poisson Distribution
- **Continuous Probability Distributions**
 - Applicable when the random variable can take on any value in a continuous range.
 - Examples:
 - Normal Distribution

4.3.1 Bernoulli Distribution

- The Bernoulli distribution is a discrete probability distribution that models the outcomes of a single trial with exactly two possible outcomes: success (often denoted by 1) and failure (often denoted by 0).
- It is one of the simplest probability distributions and is fundamental in probability theory.
- We denote the probability of success by p and the probability of failure by q where $q = 1 - p$
- Let X be a discrete random variable. Then, $X \sim \text{Bern}(p)$ means X takes on a Bernoulli distribution where the probability of success is p .

4.3.1 Key Characteristics of Bernoulli Distribution

- Single Trial: It describes the probability of success or failure in a single experiment.
- Outcomes: There are only two possible outcomes:
 - Success ($X = 1$) with probability p
 - Failure ($X = 0$) with probability q where $q=1-p$
- Parameter:
 - p : The probability of success (where $0 \leq p \leq 1$).
- Probability Mass Function (PMF): The probability mass function is given by:

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Alternatively, it can be expressed as:

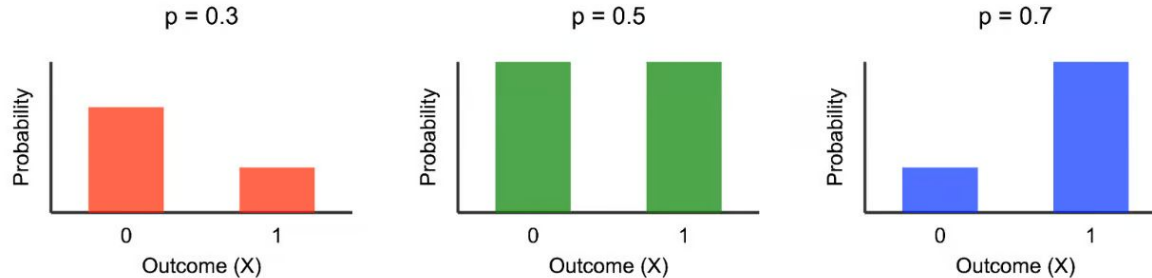
$$P(X = x) = p^x(1 - p)^{1-x}, \quad x \in \{0, 1\}$$

4.3.1 Key Characteristics of Bernoulli Distribution_[contd.]

- Mean and Variance:
 - Mean (Expected Value): $\mu = E[X] = p$
 - Variance: $\sigma^2 = \text{Var}(X) = p(1-p) = pq$ where, $q = (1-p)$
- Graph: The graph of a Bernoulli Distribution is a simple bar chart with only two bars

Comparison of Bernoulli Distributions

Probability Mass Functions for different p values



4.3.1 Bernoulli Distribution Example

- Example: Find the probability of getting heads (success) on flipping a fair coin.
- Solution:
- Let X represent the outcome of the coin toss.
- $X = 1$ if heads, $X = 0$ if tails.
- p (probability of success) = $P(X=1) = 0.5$ for a fair coin
- q (probability of failure) = $P(X=0) = q = 1 - p = 0.5$

4.3.1 Bernoulli Distribution Application

- Binary Events:
 - Success/Failure (e.g., a manufacturing defect: defective or non-defective)
 - Yes/No (e.g., customer buying a product or not)
- Foundation for Binomial Distribution:
 - The Bernoulli distribution is the building block of the Binomial distribution, which models the number of successes in n independent Bernoulli trials.
- Machine Learning and Statistics:
 - Used in logistic regression, Naïve Bayes classifiers, and other models where binary outcomes are predicted.
- Hypothesis Testing:
 - Bernoulli distribution is used in testing hypotheses about proportions in a population.

4.3.2 Binomial Distribution

- The Binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials.
- It is widely used when there are multiple trials of a binary experiment.
- Let X be the discrete random variable which counts the number of successes obtained from n Bernoulli trials. Then we denote $\mathbf{X \sim Bin(n, p)}$ which means X follows the binomial distribution with n trials, each of which have a probability \mathbf{p} of success with following probability mass function:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n$$

- $\binom{n}{k}$ is the binomial coefficient, which represents the number of ways to choose k successes from n trials:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

4.3.2 Key Characteristics of Binomial Distribution

- Multiple Trials:
 - It represents the total number of successes (X) in n independent trials.
- Outcomes per Trial:
 - Each trial has only two possible outcomes: success ($X=1$) and failure ($X=0$)
- Parameters:
 - n : The number of trials.
 - p : The probability of success in a single trial ($0 \leq p \leq 1$)
- Probability Mass Function (PMF): The probability of observing exactly k successes in n trials is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n$$

- $\binom{n}{k}$ is the binomial coefficient, which represents the number of ways to choose k successes from n trials:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

4.3.2 Key Characteristics of Binomial Distribution

- Mean, Variance, and Standard Deviation:

- **Mean (Expected Value):**

$$\mu = \mathbb{E}[X] = n \cdot p$$

- **Variance:**

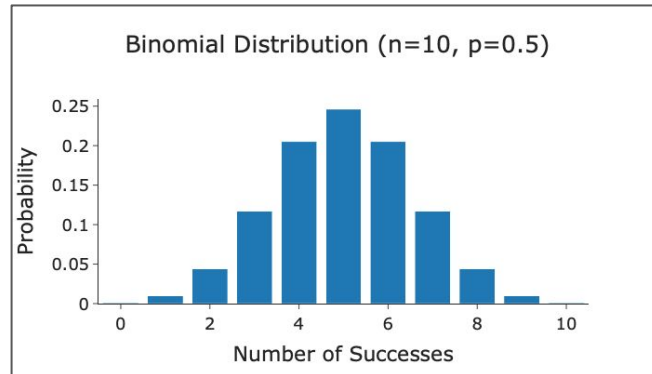
$$\sigma^2 = \text{Var}(X) = n \cdot p \cdot (1 - p)$$

- **Standard Deviation:**

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

Real World Scenarios and
Graph visualization will be
done in class

- Graph:



4.3.2 Binomial Distribution Example

- Example: You toss a coin 5 times, and you are interested in finding the probability of getting exactly 3 heads.

Solution:

- Number of trials: $n=5$ (the number of tosses).
- Probability of success (getting heads): $p=0.5$ (assuming a fair coin).
- Number of successes: $k=3$ (the number of heads we want).
- Substituting those values in PMF we get, $P(X = 3) = \binom{5}{3}(0.5)^3(1 - 0.5)^{5-3}$
- $P(X=3)=10 \times 0.125 \times 0.25 = 0.3125$
- **The probability of getting exactly 3 heads in 5 coin tosses is 0.3125 (or 31.25%).**

4.3.2 Application of Binomial Distribution

- Inspecting a batch of items to find the number of defective products.
- Modeling the success or failure of a medical treatment.
- Estimating the success rate of marketing campaigns.
- Predicting the number of people agreeing with a statement in a survey.
- Analyzing multiple-choice test performance.

4.3.3 Poisson Distribution

- The Poisson distribution is a type of discrete probability distribution that calculates the likelihood of a certain number of events happening in a fixed time or space, assuming the events occur independently and at a constant rate.
- The Poisson distribution is widely used for modeling the number of events in fixed intervals.

4.3.3 Key Characteristics of Poisson Distribution

- Discrete Nature
 - The Poisson distribution is a discrete probability distribution.
 - It models the count of events ($k=0,1,2,\dots$) in a fixed interval
- Parameter
 - λ : The average number of events in the interval (also called the rate parameter).
 - λ must be positive ($\lambda>0$).
- Probability Mass Function (PMF)
 - The PMF of the Poisson distribution gives the probability of exactly k events occurring in a fixed interval:

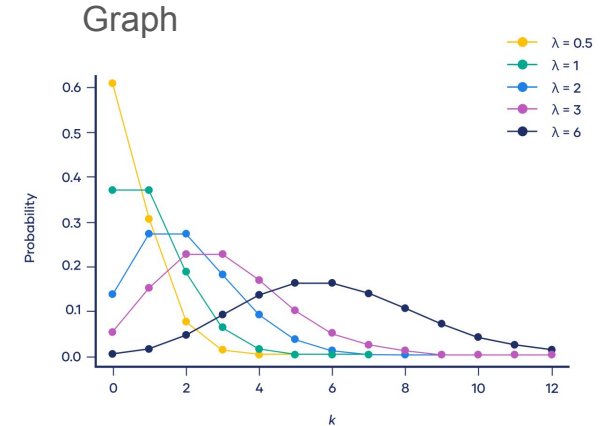
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

Where:

- X : Random variable representing the number of events.
- λ : Mean number of events per interval (rate parameter).
- e : Euler's number (≈ 2.718).
- $k!$: Factorial of k .

4.3.3 Key Characteristics of Poisson Distribution

- Independence of Events
 - Events occur independently of each other.
 - The occurrence of one event does not affect the probability of another.
- Constant Rate
 - The average rate (λ) of occurrence remains constant over the interval.
 - Example: If $\lambda=4$, then 4 events are expected on average in every unit of time or space.
- Mean and Variance
 - The mean (μ) and variance (σ^2) of the Poisson distribution are both equal to λ i.e. $\mu=\lambda, \sigma^2=\lambda$
- Skewness
 - For small λ , the distribution is positively skewed (right-skewed).
- Rare Events
 - The Poisson distribution is often used to model rare events occurring in a fixed interval



4.3.3 Poisson Distribution: Traffic Example

- Cars pass a checkpoint at an average rate of 2 cars per minute ($\lambda=2$).
- What is the probability of exactly 3 cars passing in a minute?

$$P(X = 3) = \frac{2^3 e^{-2}}{3!} = \frac{8 \cdot 0.1353}{6} \approx 0.180$$

4.3.3 Applications of Poisson Distribution

- Telecommunications: Number of calls arriving at a call center in an hour.
- Healthcare: Number of patients arriving in an emergency room per hour.
- Traffic: Number of cars passing through a checkpoint per minute.
- Natural Sciences: Number of meteors visible in an hour.

4.3.4 Normal or Gaussian Distribution

- Normal Distribution is the most common or normal form of distribution of Random Variables, hence the name “normal distribution.”
- It is also called Gaussian Distribution in Statistics or Probability.
- Normal Distribution is defined by the probability density function for a **continuous random variable** in a system.
- A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics.

4.3.4 Key Characteristics of Gaussian Distribution

- Continuous Nature
 - Gaussian Distribution is continuous probability distribution.
- Parameters:
 - Mean (μ): Determines the location of the center.
 - Standard deviation (σ): Determines the spread.
- Probability Density Function
 - The probability density function of the normal distribution is given by:

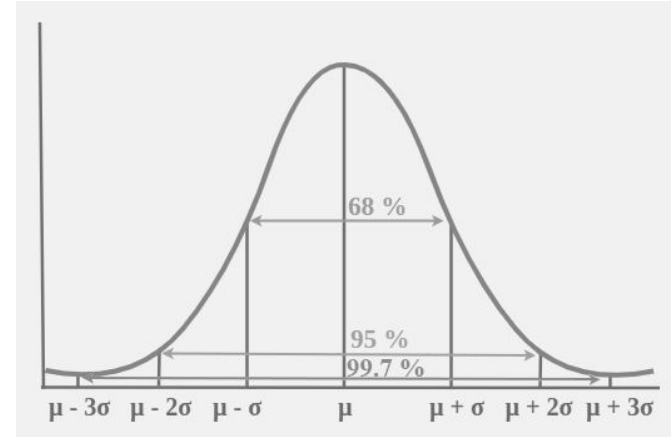
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- x : Random variable.
- μ : Mean (center of the distribution).
- σ^2 : Variance (spread of the distribution).
- σ : Standard deviation (square root of variance).

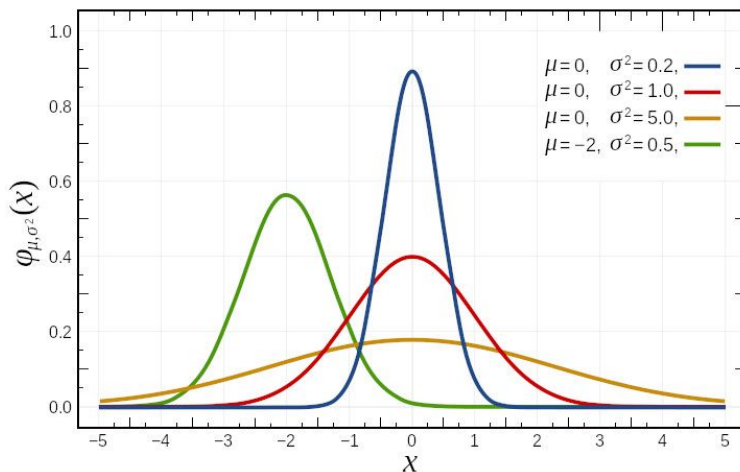
4.3.4 Key Characteristics of Gaussian Distribution

- Symmetry:
 - The distribution is perfectly symmetric about its mean (μ).
 - This implies that the left and right sides of the curve are mirror images.
- Mean, Median, and Mode:
 - All three measures are equal and located at the center of the distribution.
- Graph
 - The graph is a bell-shaped curve, symmetric around the mean (μ).
 - The curve is highest at $x = \mu$ and decreases as x moves away from the mean.
 - The curve approaches the x-axis but never touches it.
- Total Area Under the Curve:
 - The total area under the curve is 1, representing the total probability.



4.3.4 Key Characteristics of Gaussian Distribution

- Spread:
 - The spread of the curve depends on the standard deviation (σ):
 - A smaller σ results in a narrow, peaked curve.
 - A larger σ results in a wider, flatter curve.



4.3.5 Standard Normal Distribution

- The Standard Normal Distribution is a special case of the normal distribution where the **mean (μ) is 0**, and the **standard deviation (σ) is 1**.
- It serves as a reference for comparing and interpreting scores from any normal distribution.
- The random variable is called a **z-score**. Any normal distribution can be converted into a standard normal distribution using z-scores.

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

4.3.5 Standard Normal Distribution

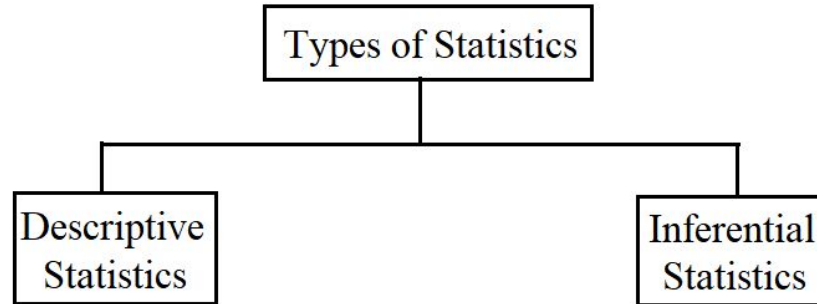
- Probability Density Function (PDF):
 - The PDF formula for the standard normal distribution is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

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

5.0 Statistics

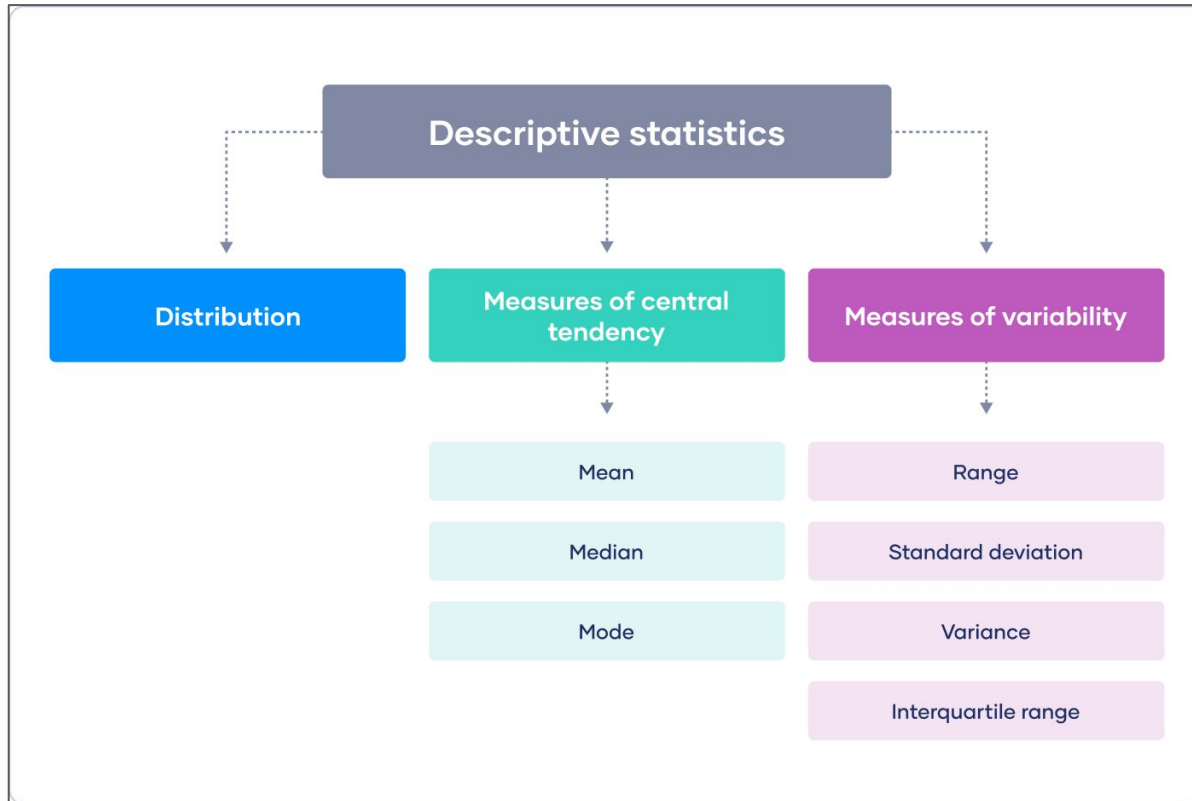
- Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, and presentation of numerical data.
- Statistics is broadly classified into **Descriptive Statistics** and **Inferential Statistics**, both of which serve different purposes in data analysis.



5.1 Descriptive Statistics

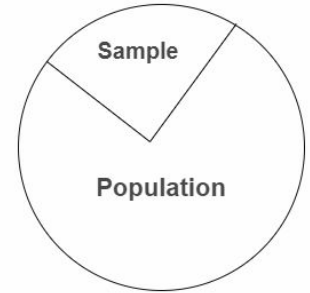
- Descriptive statistics is a branch of statistics that is used to summarize, organize, and present data in a meaningful way.
- It focuses on the "what" of the data, providing a clear overview of its characteristics.
- **Properties**
 - It describes the **dataset as it is** without making predictions or drawing conclusions beyond the data.
 - **Summarization**: Reduces large amounts of data into a concise form.Examples: Mean, median, mode
 - **Static Analysis**: Results are fixed and pertain only to the specific dataset being analyzed.
- **Tools**
 - Measures of Central Tendency: Mean, median, mode.
 - Measures of Dispersion: Range, variance, standard deviation, Interquartile Range.
 - Graphical Representations: Histograms, bar charts, pie charts, box plots.
- **Example**: Calculating the average age of students in a class.

5.1 Descriptive Statistics



5.2 Inferential Statistics

- Inferential statistics is a branch of statistics that is used to make predictions, inferences, or generalizations about a population based on a sample of data.
- It focuses on the "why" and "how" to draw conclusions about the larger context.
- **Properties**
 - Analyzes a sample to draw conclusions about the entire population.
 - It facilitates to test hypotheses, and make decisions about population parameters based on sample data.
 - **Dynamic Analysis:** Results are probabilistic and may vary depending on the sample.
- **Tools**
 - Hypothesis Testing: Determines whether a result is statistically significant.
 - Regression Analysis: Examines relationships between variables.
 - ANOVA (Analysis of Variance): Compares means across multiple groups.
- Example: Predicting the average age of all students in a school based on a random sample.



5.3 Descriptive Vs. Inferential Statistics

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Describes and summarizes data.	Makes predictions and generalizations about a population.
Scope	Focused on the sample or dataset at hand.	Generalizes findings to a larger population.
Techniques	Central tendency, dispersion, graphs.	Hypothesis testing, confidence intervals, regression.
Data Dependency	Results apply only to the dataset analyzed.	Results infer information about a larger group.
Probability Use	Not required.	Relies heavily on probability theory.
Dynamic or Static	Static (fixed for a dataset).	Dynamic (varies with sample data).

6.1 Population

- A population is the entire group of individuals, items, or events that one is interested in studying.
- It includes all possible observations relevant to a particular research question or study.
- Represents the whole group.
- Parameters (μ, σ, P) are used to describe the population (e.g., population mean μ , population proportion P).
- Typically impractical to measure directly due to size, cost, or accessibility.
- Examples:
 - All students in a university.
 - All trees in a forest.
 - Every smartphone sold in a specific year.

6.2 Sample

- A sample is a subset of the population selected for analysis.
- It is used to make inferences about the population because it is often impractical to study the entire population.
- Sample represents part of the population.
- Statistics (\bar{x} , s,p) are used to describe the sample (e.g., sample mean \bar{x} , sample proportion p).
- Sample should be chosen randomly to reduce bias and ensure representativeness.
- It is cost-effective and quicker to study.
- Example:
 - 100 students selected from a university.
 - 50 trees sampled from a forest.
 - Data from 1,000 smartphones surveyed out of millions sold.

6.3 Parameter and Point Estimate

- A parameter is a numerical value that describes a characteristic of a population. It is a fixed value but is usually unknown because measuring an entire population is impractical. Example: mean (μ) and standard deviation (σ)
- A point estimate is a single value statistic used to estimate an unknown population parameter based on sample data.
- Point estimate represents the "best guess" of the population parameter.
- Common Point Estimates are sample mean (\bar{x}), sample variance (s^2), sample proportion (p)
- Example:
 - Estimating the population mean (μ) using the sample mean (\bar{x}).
 - If the sample mean of students' heights is 165 cm, then 165 cm is the point estimate for the population mean.

6.4 Sampling Distribution

- A sampling distribution is the probability distribution of a statistic (e.g., mean, proportion, variance) computed from multiple samples of the same size drawn from the same population.
- It describes how a statistic (like the sample mean) varies across different samples.
- Sampling Distribution foundation of Inferential Statistics.
- It helps make inferences about population parameters using sample statistics.
- It is used in hypothesis testing and confidence intervals.

Example:

Consider a population of 10,000 students with a mean height (μ) of 165 cm and a standard deviation (σ) of 10 cm.

- **Step 1:** Randomly draw 100 samples, each containing 50 students.
- **Step 2:** Compute the mean height (\bar{x}) for each sample.
- **Step 3:** Plot the means ($\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{100}$).

The distribution of these sample means forms the **sampling distribution of the sample mean**.

6.5 Independent and Identically Distributed (i.i.d.)

- Random variables X_1, X_2, \dots, X_n are said to be independent if the outcome of one variable does not influence the outcomes of the others.
- Random variables X_1, X_2, \dots, X_n are identically distributed if they have the same probability distribution.
- A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent.
- Example: Measuring the heights of randomly chosen people from a population, where each measurement is independent, and the population follows the same height distribution.

7.1 Central Limit Theorem (CLT)

- CLT states that:

“For any population with a finite mean (μ) and finite variance (σ^2), the sampling distribution of the sample mean approaches a normal distribution (bell-shaped curve) as the sample size (n) becomes large, regardless of the population's original distribution.”

- The parameters of the sampling distribution of the mean are determined by the parameters of the population as follows:
 - The mean of the sampling distribution is equal to the mean of the population i.e. $\mu_{\bar{x}} = \mu$
 - The standard deviation of the sampling distribution also known as **standard error** is equal to the standard deviation of the population divided by the square root of the sample size i.e.

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

7.1 Central Limit Theorem (CLT)

We can describe the sampling distribution of the mean using this notation:

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

Where:

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

7.1 Conditions/Assumptions for CLT

- The sample size must be sufficiently large. This condition is usually met if the sample size is $n \geq 30$.
- The samples must be independent and identically distributed (i.i.d.) random variables.
- CLT only holds for a population with finite variance.
- When the sampling is done without replacement, the sample size shouldn't exceed 10% of the total population.

7.1 Normal Approximation to Binomial Distribution

- Explanation and Numerical will be performed in class.

7.1 Normal Approximation to Poisson Distribution

- Explanation and Numerical will be performed in class.

7.1 Numericals

1. A factory produces 90% of its products without defects. If one product is selected at random, what is the probability it is not defective?
2. A coin is flipped 10 times. The probability of getting heads in each flip is 0.5.
 - a. What is the probability of getting exactly 6 heads?
 - b. What is the probability of getting at least 8 heads?
3. A call center receives an average of 5 calls per hour.
 - a. What is the probability that the call center receives exactly 7 calls in an hour?
 - b. What is the probability of receiving fewer than 3 calls in an hour?
4. A marketing campaign has a 70% success rate for convincing customers to subscribe to a service. A customer is chosen at random.
 - a. What is the probability the customer subscribes?
 - b. What is the probability the customer does not subscribe?
5. A website has a 5% conversion rate (i.e., 5% of visitors make a purchase). If 20 visitors visit the website:
 - a. What is the probability exactly 2 visitors make a purchase?
 - b. What is the probability at least 1 visitor makes a purchase?
6. In a spam detection system, an email is flagged as spam with a probability of 0.8.
 - a. If one email is chosen at random, what is the probability it is flagged as spam?
 - b. What is the probability it is not flagged as spam?

7.1 Numericals

1. An online store receives an average of 10 orders per hour.
 - a. What is the probability the store receives exactly 15 orders in an hour?
 - b. What is the probability it receives fewer than 5 orders in an hour?
2. A quality control inspector tests 10 items from a batch where 2% of the items are defective.
 - a. What is the probability exactly 1 item is defective?
 - b. What is the probability no items are defective?
3. A quadratic cost function is given as: $J(\theta) = (\theta - 5)^2$. Using gradient descent: If the initial value of θ is 0 and the learning rate (α) is 0.1, compute the first 5 iterations of θ . What is the optimal value of θ ?

7.1 Numericals

Consider a dataset with two features, x_1 and x_2 , and the hypothesis function:

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

with the cost function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2$$

For the following dataset:

$$(x_1, x_2, y) = \{(1, 2, 4), (2, 1, 4.5), (3, 3, 6)\}$$

1. Derive the gradient descent update rules for $\theta_0, \theta_1, \theta_2$.
2. Perform 2 iterations of gradient descent starting with $\theta_0 = 0, \theta_1 = 0, \theta_2 = 0$, and $\alpha = 0.05$.

8.1 Hypothesis Testing: Terminologies

- **Hypothesis**

- Hypothesis is an assumption specifically a statistical claim about an unknown population parameter.

- **Hypothesis testing**

- Hypothesis testing is a statistical method used to decide whether there is enough evidence in a sample of data to support a particular claim about a population.

- **Null Hypothesis (H_0)**

- The null hypothesis is a statement that assumes no effect, no difference, or no relationship in the population. It represents the default assumption that any observed changes are due to random chance.

- **Alternative Hypothesis (H_1 or H_a)**

- The alternative hypothesis contradicts the null hypothesis and represents the claim we are testing. It suggests the presence of an effect, difference, or relationship.

8.1 Hypothesis Testing: Terminologies

- **p-value**

- A p-value (probability value) is a measure that helps determine the significance of the results in a hypothesis test.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- A small p-value suggests strong evidence against H_0 .
- A large p-value suggests weak evidence against H_0 .
- The p-value is compared to the significance level (α).
- P-values are usually calculated using statistical software or p-value tables based on the assumed or known probability distribution of the specific statistic tested.

- **Significance Level or Level of Significance (α)**

- The significance level (α) is the probability of rejecting the null hypothesis (H_0) when it is actually true.
- It represents the risk of making a Type I Error (false positive).
- It is chosen before conducting the test.
- The most common significance level is 0.05 (5%), meaning there is a 5% chance of rejecting H_0 when it is true.
- If $p\text{-value} < \alpha \rightarrow$ Reject H_0 (The result is statistically significant).
- If $p\text{-value} \geq \alpha \rightarrow$ Fail to reject H_0 (Not enough evidence to support H_1).





8.1 Hypothesis Testing: Terminologies

- **Type I Error (False Positives)**

- A Type I Error occurs when we reject a null hypothesis (H_0) even it is true.
- Example: The test result says you have coronavirus, but you actually don't.

- **Type II Error (False Negatives)**

- A Type II Error occurs when we fail to reject a null hypothesis (H_0) when it is false.
- Example: The test result says you don't have coronavirus, but you actually do.

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	 Type I Error (False positive)	 Correct Outcome! (True positive)
Fail to reject null hypothesis	 Correct Outcome! (True negative)	 Type II Error (False negative)

8.2 Hypothesis Testing Procedures

1. State Null Hypothesis(H_0)
2. State Alternative Hypothesis(H_1 or H_a)
3. Choose the Significance Level (α)
4. Select the Appropriate Test
 - Z-Test: When population variance is known, large sample size ($n > 30$).
 - T-Test: When population variance is unknown, small sample size ($n \leq 30$).
 - ANOVA (Analysis of Variance): Comparing means of three or more groups.
5. Calculate the Test Statistic
6. Find the Critical Value
7. Compare the test statistic with the critical value
8. Make a decision and interpret the results

8.3 One Sample Z-Test: Mean of a population

- The one-sample Z-test is used to determine whether the mean of a single sample is significantly different from a known population mean when the population standard deviation (σ) is known and the sample size is large ($n > 30$).

8.3 One Sample z-Test: Procedures

1. State H_0 (Null Hypothesis): The sample mean is equal to the population mean. $H_0 : \mu = \mu_0$
2. State H_1 (Alternative Hypothesis): The sample mean is different from (\neq), greater than ($>$), or less than ($<$) the population mean.
3. Choose the Significance Level (α)
4. Z-Test Statistic:

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

Where:

- \bar{X} = Sample mean
- μ = Population mean
- σ = Population standard deviation (known)
- n = Sample size

5. Find the Critical Value from Z-table
6. Compare the test statistic with the critical value
 - Reject H_0 if $|Z\text{-calculated}| > |Z\text{-critical}|$ else accept H_0
7. Make a decision and interpret the results

8.3 One Sample t-Test: Mean of a population

- The one-sample t-test is used to determine whether the mean of a single sample is significantly different from a known population mean when the population standard deviation (σ) is unknown and the sample size is small ($n \leq 30$).

8.3 One Sample t-Test: Procedures

1. State H_0 (Null Hypothesis): The sample mean is equal to the population $H_0 : \mu = \mu_0$
2. State H_1 (Alternative Hypothesis): The sample mean is different from (\neq), greater than ($>$), or less than ($<$) the population mean.
3. Choose the Significance Level (α)
4. T-Test Statistic:

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

Where:

- \bar{X} = Sample mean
- μ = Population mean
- s = Sample standard deviation
- n = Sample size

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

5. Find the Critical Value from t-table for $df = n-1$
6. Compare the test statistic with the critical value
 - Reject H_0 if $|t\text{-calculated}| > t\text{-critical}$.
7. Make a decision and interpret the results

8.4 Two Sample Z-Test: Difference between two population means

- The two-sample Z-test is used to compare the means of two independent populations to determine whether there is a statistically significant difference between them.
- When to Use a Two-Sample Z-Test?
 - When comparing the means of two independent groups.
 - When the population variances (σ^2) are known, or the sample sizes are large ($n_1, n_2 \geq 30$).
- Test Statistics:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- \bar{X}_1, \bar{X}_2 = Sample means of the two populations
- μ_1, μ_2 = Population means (usually assumed equal under H_0)
- σ_1, σ_2 = Population standard deviations
- n_1, n_2 = Sample sizes

8.4 Two Sample t-Test: Difference between two population means

- Two Sample t-Test is used when we need to compare the statistical means of two independent samples or groups. It helps us determine whether there is a significant difference between the means of the two groups.
- Test Statistics for equal variance:

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

Where:

- \bar{X}_1, \bar{X}_2 = Sample means of the two populations
- s_1, s_2 = Sample standard deviations
- n_1, n_2 = Sample sizes
- df = $n_1 + n_2 - 2$

The **pooled variance** (s_p^2) is used when assuming equal population variances:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

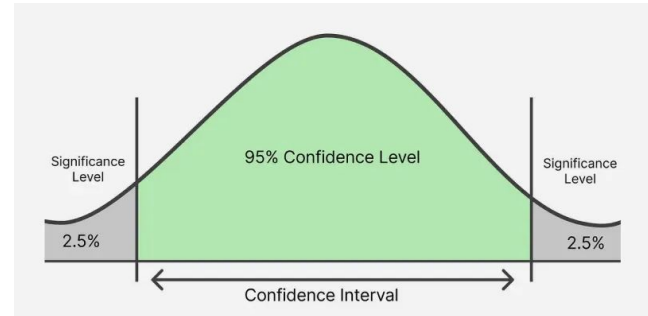
8.5 Confidence Intervals

- A Confidence Interval (CI) is a range of values used to estimate a population parameter (such as the mean or proportion) based on sample data.
- It provides an interval within which the true population parameter is likely to fall with a certain level of confidence (e.g., 95%, 99%).
- Interpreting Confidence Intervals
 - Let's say we take a sample of 50 students and calculate a 95% confidence interval for their average height which turns out to be 160–170 cm. This means If we repeatedly take similar samples 95% of those intervals would contain the true average height of all students in the population.

- Formula:

$$CI = \bar{x} \pm Z * \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$



8.1 Numericals

1. A factory claims that the average lifetime of its light bulbs is 1,200 hours. A sample of 40 light bulbs is tested, and the sample mean is found to be 1,180 hours. The standard deviation of the lifetime of the light bulbs is known to be 100 hours. Conduct a one-sample z-test at the 5% significance level to determine whether the sample mean is significantly different from the claimed average.
2. A researcher claims that the average IQ score of a population is 100. A sample of 50 individuals is taken, and the sample mean IQ score is found to be 102 with a population standard deviation of 15. Conduct a one-sample z-test at the 1% significance level to determine whether the sample mean is significantly greater than 100.
3. A company claims that the average salary of its employees is \$50,000. A sample of 60 employees is selected, and the sample mean salary is \$52,000 with a population standard deviation of \$8,000. Conduct a one-sample z-test at the 10% significance level to determine whether the sample mean salary is significantly different from \$50,000.
4. A pharmaceutical company claims that the average blood pressure reduction after taking their drug is 15 mmHg. A sample of 36 patients is tested, and the sample mean reduction is 13 mmHg with a population standard deviation of 4 mmHg. Conduct a one-sample z-test at the 1% significance level to determine whether the sample mean is significantly less from 15 mmHg.

8.1 Numericals

1. A group of 15 students is tested to find out the average time they spend on homework each week. The sample mean is 12 hours, and the sample standard deviation is 3 hours. Test at the 5% significance level if the average time spent on homework is different from 10 hours.
2. A school claims that the average score on their final exam is 75. A sample of 10 students' scores are taken, and the scores are as follows: 78,74,82,70,76,73,80,79,72,75. Determine whether the average score is significantly different from 75 at the 5% significance level.
3. A researcher claims that college students sleep an average of 7 hours per night. A random sample of 12 students reports their sleep durations (in hours) over a night: 6.5, 7.2, 6.8, 5.9, 7.4, 6.1, 7.0, 6.3, 6.7, 7.1, 5.8, 6.9

At a 0.01 significance level, test whether the actual average sleep duration is different from 7 hours.

8.1 Numericals

- A company wants to estimate the average salary of software engineers. A random sample of 50 engineers is taken, and their average salary is found to be \$80,000. The population standard deviation is known to be \$12,000. Find the 95% confidence interval for the true mean salary.

Therefore, we are 95% confident that the true average salary of software engineers is between \$76,672 and \$83,328.

1. Given Data:

- Sample Mean $\bar{X} = 80,000$
- Population Standard Deviation $\sigma = 12,000$
- Sample Size $n = 50$
- Confidence Level = **95%**
(From the Z-table, **Z-value for 95% CI = 1.96**)

2. Confidence Interval Formula (Z-Test):

$$CI = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

3. Calculate the Standard Error (SE):

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{12,000}{\sqrt{50}} = \frac{12,000}{7.07} = 1,697.06$$

4. Calculate the Margin of Error (ME):

$$ME = 1.96 \times 1,697.06 = 3,328.24$$

5. Confidence Interval:

$$80,000 \pm 3,328.24$$

$$(76,671.76, 83,328.24)$$

8.1 Numericals

- A researcher wants to estimate the average height of male students in a university. A random sample of 10 students is selected, and their average height is 175 cm with a sample standard deviation of 8 cm. Find the 95% confidence interval for the true mean height.

Therefore, we are 95% confident that the true average height of male students is between 169.28 cm and 180.72 cm.

1. Given Data:

- Sample Mean $\bar{X} = 175$
- Sample Standard Deviation $s = 8$
- Sample Size $n = 10$
- Confidence Level = **95%**
(From the t-table, **t-value for $df = 9$ at 95% = 2.262**)

2. Confidence Interval Formula (T-Test):

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

3. Calculate the Standard Error (SE):

$$SE = \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{10}} = \frac{8}{3.16} = 2.53$$

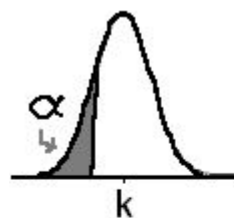
4. Calculate the Margin of Error (ME):

$$ME = 2.262 \times 2.53 = 5.72$$

5. Confidence Interval:

$$175 \pm 5.72$$
$$(169.28, 180.72)$$

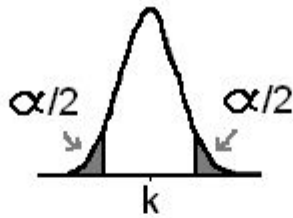
8.1 Numericals



$$H_0: \mu = k$$

$$H_1: \mu < k$$

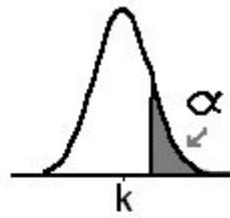
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33



$$H_0: \mu = k$$

$$H_1: \mu \neq k$$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58



$$H_0: \mu = k$$

$$H_1: \mu > k$$

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

Degrees of Freedom (df)	$\alpha = 0.05$	$\alpha = 0.01$
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845

9.1 ANOVA (Analysis of Variance)

- Analysis of Variance (ANOVA) is a statistical method used to compare the means of three or more groups to determine if there is a significant difference among them.
- ANOVA tests the null hypothesis that all group means are equal, against the alternative hypothesis that at least one group mean is different.
- Instead of performing multiple t-tests, which increases the probability of making a Type I error, ANOVA allows us to test all groups simultaneously with a single hypothesis test.
- Assumptions for ANOVA
 - The dependent variable is approximately normally distributed within each group.
 - The samples are selected at random and should be independent of one another.
 - All groups have equal standard deviations.
 - Each data point should belong to one and only one group. There should be no overlap or sharing of data points between groups.

9.1 Types of ANOVA

- One-Way ANOVA

- This is the most basic form of ANOVA and is used when there is only one independent variable with more than two levels or groups. It assesses whether there are any statistically significant differences among the means of the groups.
- It compares means across one independent variable.
- Example: Comparing the average test scores of students from three different schools.

- Two-Way ANOVA

- It compares means across two independent variables.
- It allows for the examination of the main effects of each variable as well as the interaction between them. The interaction effect explores whether the effect of one variable on the dependent variable is different depending on the level of the other variable.
- Example: Analyzing the effect of diet type and exercise level on weight loss.

9.2 One-Way ANOVA: Procedures

- A teacher wants to compare the average test scores of students taught using three different teaching methods. perform One-Way ANOVA at $\alpha=0.05$ to determine if there is a significant difference between the mean test scores.

Teaching Method	Scores
Method A	85, 90, 78, 92, 88
Method B	75, 80, 85, 70, 90
Method C	65, 70, 72, 68, 66

- State H_0 (Null Hypothesis): There is no difference in the mean scores across the three teaching methods. $\mu_A = \mu_B = \mu_C$
- State H_1 (Alternative Hypothesis): At least one group mean is significantly different from the others.
- The Significance Level (α) = 0.05

9.2 One-Way ANOVA: Procedures

4. Compute Group means:

$$\bar{X}_A = \frac{85 + 90 + 78 + 92 + 88}{5} = 86.6$$

$$\bar{X}_B = \frac{75 + 80 + 85 + 70 + 90}{5} = 80.0$$

$$\bar{X}_C = \frac{65 + 70 + 72 + 68 + 66}{5} = 68.2$$

$$\bar{X}_{\text{overall}} = (86.6 + 80.0 + 68.2) / 3 = 78.27$$

5. Compute Sum of Squares:

a. Sum of Squares Between Groups (SSB)

$$SSB = n_A(\bar{X}_A - \bar{X}_{\text{overall}})^2 + n_B(\bar{X}_B - \bar{X}_{\text{overall}})^2 + n_C(\bar{X}_C - \bar{X}_{\text{overall}})^2$$

$$SSB = 5(86.6 - 78.27)^2 + 5(80 - 78.27)^2 + 5(68.2 - 78.27)^2$$

$$SSB = 5(8.33)^2 + 5(1.73)^2 + 5(10.07)^2$$

$$SSB = 346.35 + 14.9 + 508.15 = 869.4$$

9.2 One-Way ANOVA: Procedures

5. Compute Sum of Squares:

b. Sum of Squares Within Groups (SSW)

$$SSW = \sum (X - \bar{X})^2 \text{ for each group}$$

For **Method A**:

$$(85 - 86.6)^2 + (90 - 86.6)^2 + (78 - 86.6)^2 + (92 - 86.6)^2 + (88 - 86.6)^2 = 29.8$$

For **Method B**:

$$(75 - 80)^2 + (80 - 80)^2 + (85 - 80)^2 + (70 - 80)^2 + (90 - 80)^2 = 62.5$$

For **Method C**:

$$(65 - 68.2)^2 + (70 - 68.2)^2 + (72 - 68.2)^2 + (68 - 68.2)^2 + (66 - 68.2)^2 = 8.2$$

$$\mathbf{SSW=29.8+62.5+8.2=100.5}$$

9.2 One-Way ANOVA: Procedures

6. Compute Degrees of Freedom

- Degrees of freedom between groups:

$$df_{\text{between}} = k - 1 = 3 - 1 = 2$$

- Degrees of freedom within groups:

$$df_{\text{within}} = N - k = 15 - 3 = 12$$

- Total degrees of freedom:

$$df_{\text{total}} = df_{\text{between}} + df_{\text{within}} = 2 + 12 = 14$$

7. Compute Mean Squares

- Mean Square Between Groups (MSB):

$$MSB = \frac{SSB}{df_{\text{between}}} = \frac{869.4}{2} = 434.7$$

- Mean Square Within Groups (MSW):

$$MSW = \frac{SSW}{df_{\text{within}}} = \frac{100.5}{12} = 8.375$$

9.2 One-Way ANOVA: Procedures

8. Compute F-Statistics:

$$F = \frac{MSB}{MSW} = \frac{434.7}{8.375} = 51.9$$

9. Compute Critical Value

Using an **F-table** at $\alpha = 0.05$, with $df_{\text{between}} = 2$ and $df_{\text{within}} = 12$, the **critical F-value** is **3.89**.

10. Compare F calculated and F-table

$$F_{\text{calc}} = 51.9 > 3.89 = F_{\text{critical}}$$

Since F-calculated > F-critical, we reject H₀

11. Conclusion:

At least one teaching method leads to significantly different test scores.

9.2 F-Table

F - Distribution ($\alpha = 0.05$ in the Right Tail)

Denominator Degrees of Freedom df_2	df_1	Numerator Degrees of Freedom								
		1	2	3	4	5	6	7	8	9
1	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	1	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	1	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	1	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	1	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	1	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	1	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	1	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	1	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	1	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	1	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	1	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	1	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	1	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	1	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	1	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	1	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	1	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	1	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	1	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	1	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	1	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	1	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	1	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	1	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	1	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	1	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501

10.0 Bayes' Theorem

- Bayes theorem is a mathematical formula that calculates the conditional probability of event A given the occurrence of event B.

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

- $P(A|B)$: Posterior Probability
 - The probability of event A (hypothesis) being true, given the observed evidence B.
- $P(B|A)$: Likelihood
 - The probability of observing evidence B, assuming A is true.
- $P(A)$: Prior Probability
 - The probability of A being true before observing any evidence.
- $P(B)$: Evidence (Normalization Factor)
 - The probability of observing evidence B under all possible scenarios.

10.0 Bayes' Theorem: Numerical

- A company manufactures 40% of its products in Factory A, 35% in Factory B, and 25% in Factory C. The probability that a product is defective is:
 - Factory A: 5% defective
 - Factory B: 4% defective
 - Factory C: 2% defective
- If a randomly selected product is found to be defective, what is the probability that it was manufactured in Factory B?