UNIT 2

Mathematics for Data Science

2.1 Introduction to linear algebra for data science

2.2 Vectors, matrices and matrix factorization

2.3 Gradient descent for optimization

2.4 Introduction to probability and random variable

2.5 Probability distributions: Normal, Bernoulli, Binomial, Poisson

2.6 Descriptive and inferential statistics

2.7 Central limit theorem and sample distribution concepts

2.8 Normal approximation; hypothesis testing procedures: Tests about the mean of a normal population

2.9 The t-test, Z-tests for differences between two populations means, the two-sample t-test, confidence interval for mean of normal population

2.10 ANOVA

2.1 Introduction to linear algebra for data science

Linear algebra is a branch of math that's more about vectors and matrices than about basic equations like y = mx + b . Many people think it's about plotting lines, but it's actually much broader and more abstract. A better name might be "vector algebra" or "matrix algebra" because those are its main focus.

Linear algebra is all about understanding systems of equations in terms of vector spaces and matrices. Don't worry if you're not sure what vectors or matrices are yet; we'll cover those concepts in detail. Linear algebra is a core part of many fields, like math, statistics, data science, and machine learning. Even if you don't realize it, when you work with data in these fields, you're probably using linear algebra.

What we learned ? : Linear algebra is the study of vectors and certain algebra rules to manipulate vectors.

What Is a Vector?

A vector is essentially an arrow in space that has a specific direction and length, often used to represent data. It is a fundamental component of linear algebra, forming the basis for concepts like matrices and linear transformations. In its fundamental form, it has no concept of location so always imagine its tail starts at the origin of a Cartesian plane (0,0).

Note: Vectors are usually represented as bold lower-case characters like **v**, **w** etc. Since writing boldface characters using pen and paper is difficult, it's also represented with an arrow on top of lower-case characters when using pen and paper. For this article, we will stick with boldface representation.

Graphically, vectors are depicted as arrows. The length of the arrow indicates the magnitude (or intensity) of the vector, while the angle it makes with a reference direction (usually the horizontal axis) shows the vector's direction.

To emphasize again, the purpose of the vector is to visually represent a piece of data. For example, if you have the heights, weights, and ages of a large number of people, you can treat your data as three-dimensional vectors [height, weight, age]. This vector holds three different bits of information about our data. This is the perfect use of a vector in data science. Also, if you're teaching a class with four exams, you can treat student grades as four-dimensional vectors [exam1, exam2, exam3, exam4]. Each color is represented by 3-vector (RGB).

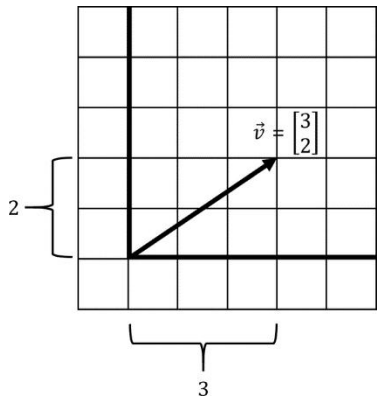
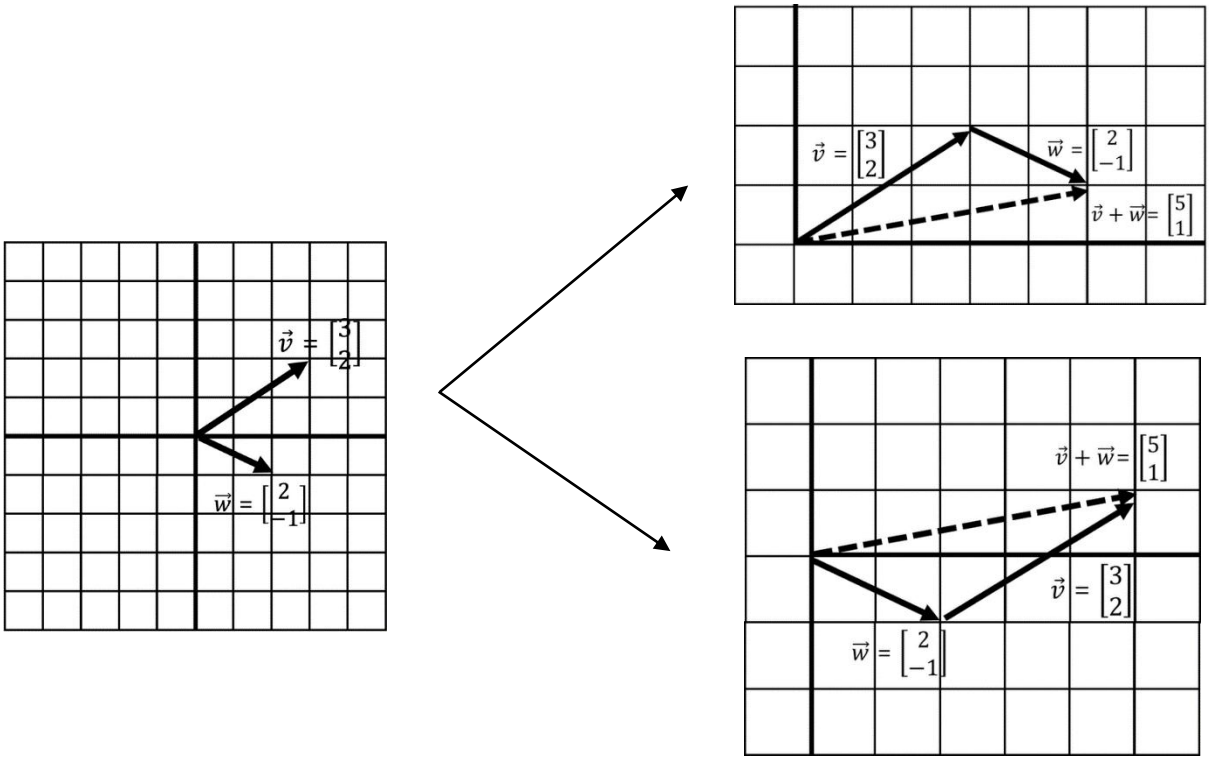


Figure shows a vector in two dimensions

Note: Vectors can exist on more than two dimensions.

Adding and Combining Vectors:

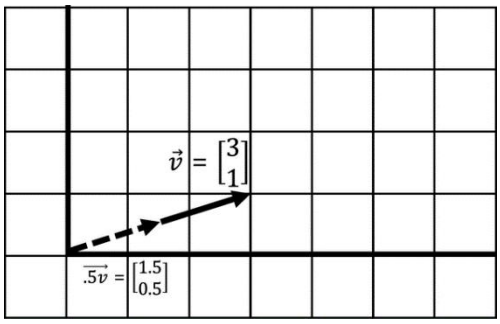


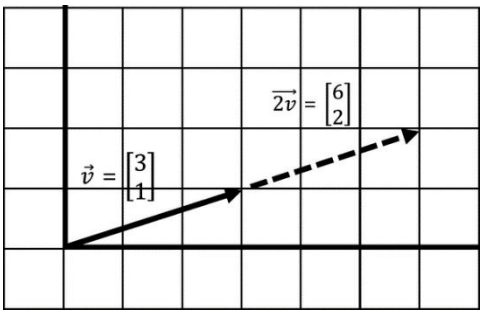
Note that it does not matter whether we add **v** and **w** before or vice versa, which means it is commutative and order of operation does not matter.

Scaling Vectors:

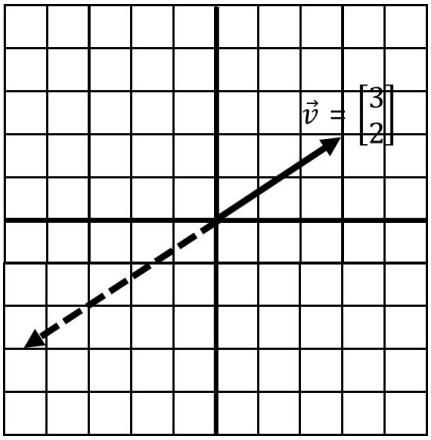
Scaling is growing or shrinking a vector's length. You can grow/shrink a vector by multiplying or scaling it with a single value, known as a scalar.

Figure 1 shows vector being scaled by a factor of 2, which doubles it. **Figure 2** shows vector being scaled down by factor of .5, which halves it.





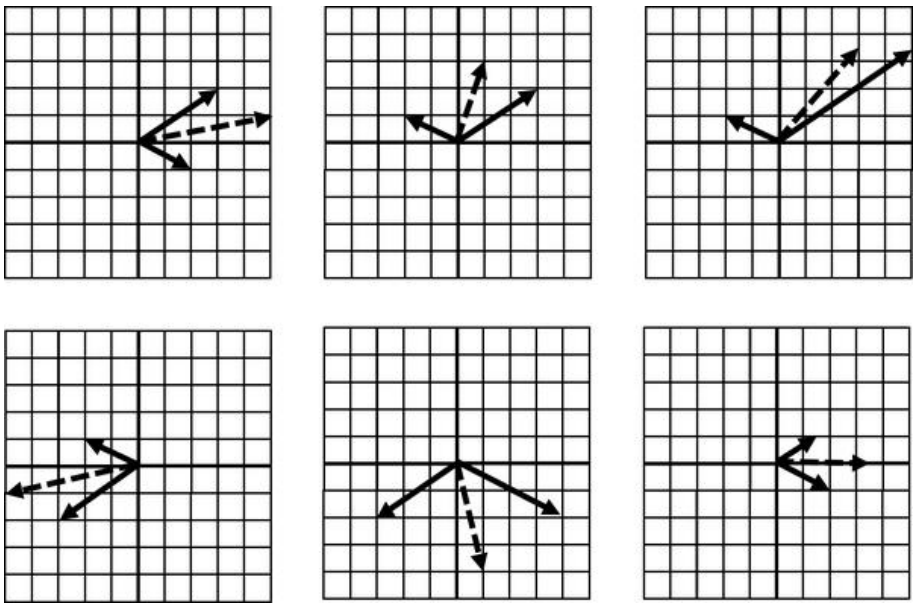
Note: An important detail to note here is that scaling a vector does not change its direction, only its magnitude. When you multiply a vector by a negative number, it flips the direction of the vector as shown in the image.



What we learned? : Manipulating data is Manipulating Vectors.

Span and Linear Independence and Dependence:

These two operations, adding two vectors and scaling them, brings about a simple but powerful idea. With these two operations, we can combine two vectors and scale them to create any resulting vector we want. Figure below shows six examples of taking two vectors ***v*** and ***w*** , and scaling and combining. These vectors fixed in two different directions, can be scaled and added to create any new vector ***v+w***.



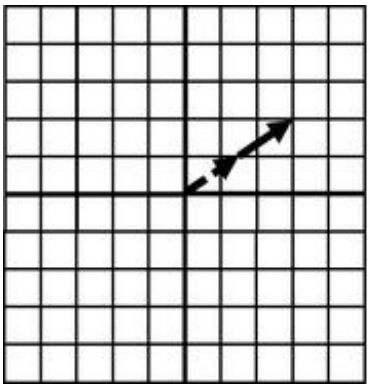
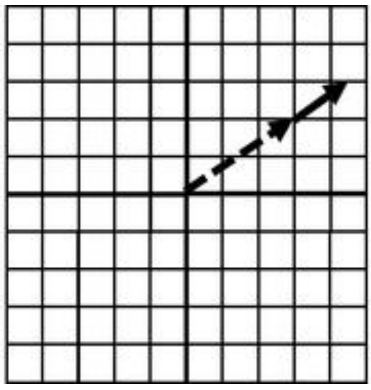
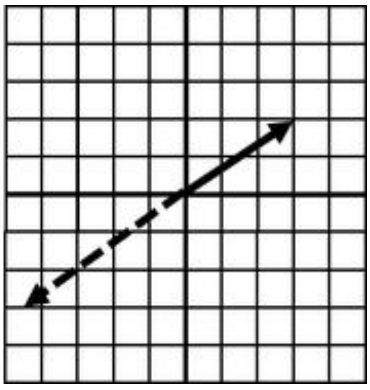
This whole space of possible vectors is called **span**, and in most cases our span can create unlimited vectors off those two vectors, simply by scaling and summing them.

When we have two vectors in two different directions, they are linearly independent and have this unlimited span.

OR,

To determine if these vectors are linearly independent, we need to check if the only solution to the equation: **$c_1v_1 + c_2v_2 = 0$ is $c_1=0$ and $c_2=0$.**

What happens when two vectors exist in the same direction, or exist on the same line? The combination of those vectors is also stuck on the same line, limiting our span to just that line. No matter how you scale it, the resulting sum vector is also stuck on that same line. This makes them linearly dependent, as shown

For example: Let v1=[1 2] and v2=[2 4]. Here, we notice that:

Since $v2$ can be expressed as a multiple of $v1$, the vectors are linearly dependent.

Types of Vectors:

Zero Vector: A vector is said to be a zero vector if the magnitude of the vector is zero. It is denoted by: O = (0, 0, 0) in 3D space or O=[0,0] in 2D space.
Unit Vector: A vector is said to be a unit vector if the magnitude of the vector is one. It is denoted by a cap.

$$\hat{\mathbf{a}} = \frac{\mathbf{a}}{\|\mathbf{a}\|}$$

where:

- $\hat{\mathbf{a}}$ represents the unit vector in the direction of $\mathbf{a}$.
- $\|\hat{\mathbf{a}}\| = 1$, which confirms that $\hat{\mathbf{a}}$ is a unit vector.

A unit vector retains the direction of the original vector a, but scales it to a length of 1.

Negative Vector: A vector is said to be a negative vector of a given vector if it has the same magnitude but points in the opposite direction.
Parallel Vectors(Collinear): Two vectors, a, and b are said to parallel if they have the same direction but not the same magnitude.

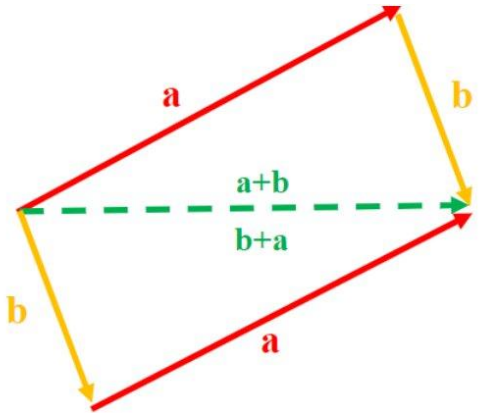
Equal Vectors: Two vectors (a and b) are said to be equal if they have the same magnitude and direction.

Orthogonal/Perpendicular Vectors: Two vectors, a, and b, are orthogonal if and only if they are perpendicular to each other. The angle between them is a right angle. Mathematically, vectors are orthogonal if two vectors are non-zero and their dot product of vectors is zero.

Operation on Vectors:

Addition of Vectors

Two vectors, a, and b, are added using the Triangle Law of Addition. If two vectors, a, and b, are represented as the side of a triangle with the magnitude and direction, then the third side of the triangle (magnitude and direction) will be the resultant vector.



If $a = a1i + b1j + c1k$, $b = a2i + b2j + c2k$, then,

$a+b = (a1+a2)i + (b1+b2)\,j + (c1+c2)k$

Vector addition follows:

Commutative law, i.e., $a + b = b + a$

Associative law, i.e., $a + (b + c) = (a + b) + c$

Subtraction of two Vectors

Vector subtraction is similar to vector addition, i.e., if a and b are two vectors, such that $a = a1i + b1j + c1k$, $b = a2i + b2j + c2k$, then,

$a - b = a + (-b) = (a1 - a2)\,i + (b1 - b2)\,j + (c1 - c2)\,k$

Product

Two vectors, a, and b, can be multiplied in two ways:

Dot Product

A dot product (or a scalar product) is a mathematical operation that takes two products and returns a scalar product. It is calculated by multiplying the corresponding elements of the vectors.

Definition of Dot Product

If a and b are two vectors, a = (a1, a2, a3, ⋯, an), and b = (b1, b2, b3, ⋯, bn), then

$a \cdot b = (a1 * b1) + (a2 * b2) + \cdots + (an * bn)$

Geometric Interpretation of the Dot Product:

If the angle between two vectors is given, then:

$a \cdot b = |a| \cdot |b| \cos(\theta)$, where, theta is the angle between a and b

$|a|$, and $|b|$ are the magnitude of a and b.

Direction of the Dot Product:

- When $\theta = 0^\circ$ (vectors point in the same direction), $\cos\theta = 1$, so the dot product is maximized, equal to $|a||b|$.
- When $\theta = 90^\circ$ (vectors are orthogonal, or perpendicular), $\cos\theta = 0$, so the dot product is zero. This indicates that orthogonal vectors have no "similarity" in direction.
- When $\theta = 180^\circ$ (vectors point in opposite directions), $\cos\theta = -1$, so the dot product is negative, equal to $-|a||b|$.

Cross Product

Cross product is a binary operation (multiplication) that is performed on two vectors, and the resultant vector is perpendicular to both the given vectors. It is calculated by using the determinants, i.e.,

Definition of the Cross Product:

If a = a1i + a2j + a3k, and b = b1i + b2j + b3k, then

$$A \times B = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$$

Expanding this determinant yields:

$$A \times B = \mathbf{i}(a_2b_3 - a_3b_2) - \mathbf{j}(a_1b_3 - a_3b_1) + \mathbf{k}(a_1b_2 - a_2b_1)$$

Geometric Interpretation of the Cross Product:

If the angle between two vectors is given, then:

In vector form, if the angle and magnitudes are known but not the components, you can express the cross product as:

$$\mathbf{A} \times \mathbf{B} = |\mathbf{A}|\,|\mathbf{B}|\,\sin\theta\,\hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is a unit vector perpendicular to both **A** and **B**.

This magnitude represents the **area of the parallelogram** formed by A and B.

Direction of the Cross Product:

- If θ is 0° or 180° (i.e., A and B are parallel or anti-parallel), sin θ= 0, so A×B=0. This means that parallel vectors have no perpendicular component.

- If $\theta=90^{\circ}$ (i.e., aa and bb are perpendicular), $\sin\theta=1$, so the magnitude of the cross product is maximized and equal to $|A||B|$.

Dot Product	Cross Product
Product of magnitude of vectors and cos of the angle between them.	Product of magnitude of vectors and sine of the angle between them.
In terms of vectors A and B $\mathbf{A \cdot B = |A||B| \cos\theta}$	In terms of vectors A and B $\mathbf{A \cdot B = |A||B| \sin\theta\, n}$
The final product is a scalar quantity.	The final product is a vector quantity.
Follows a commutative law: A.B=B.A	Does not follow a commutative law: AxB is not equal to BxA
If the vectors are perpendicular to each other, their dot result is 0. As in, A.B=0	If the vectors are parallel to each other, their cross result is 0. As in, AxB=0

Concept for Basis Vector:

Basis vectors form the foundational building blocks of a vector space. They allow every vector within that space to be uniquely represented by a linear combination of these basis vectors.

Properties of Basis Vector:

1. A set of basis vectors must **span** the vector space, meaning any vector in the space can be formed as a linear combination of these basis vectors. f you have a vector space V of dimension n, you need exactly n basis vectors to span the space.

2. The basis vectors in a vector space must be **linearly independent**. This means that no basis vector can be written as a combination of the others.

Basis Vectors that Are Unit Vectors:

Consider the vector space $\mathbb{R}^2$ (2-dimensional real space). The **standard basis vectors** are:

$$\mathbf{e}_1 = [1, 0] \quad \text{and} \quad \mathbf{e}_2 = [0, 1]$$

- These vectors are **unit vectors**, since:

$$\|\mathbf{e}_1\| = \sqrt{1^2 + 0^2} = 1 \quad \text{and} \quad \|\mathbf{e}_2\| = \sqrt{0^2 + 1^2} = 1$$

- The set $\{\mathbf{e}_1, \mathbf{e}_2\}$ is a **basis** for $\mathbb{R}^2$, because:
  - The vectors are **linearly independent** (you cannot express $\mathbf{e}_1$ as a scalar multiple of $\mathbf{e}_2$ and vice versa).
  - The vectors span $\mathbb{R}^2$, meaning any vector $\mathbf{v} = [x, y] \in \mathbb{R}^2$ can be written as:

$$\mathbf{v} = x \cdot \mathbf{e}_1 + y \cdot \mathbf{e}_2$$

Thus, the basis vectors in this case are unit vectors.

Basis Vectors that Are Not Unit Vectors:

Consider the set of vectors $\mathbf{b}_1 = [2, 0]$ and $\mathbf{b}_2 = [0, 3]$ in $\mathbb{R}^2$.

- These vectors are **linearly independent** because $\mathbf{b}_1$ cannot be written as a scalar multiple of $\mathbf{b}_2$ and vice versa.
- These vectors **span** $\mathbb{R}^2$, since any vector $\mathbf{v} = [x, y] \in \mathbb{R}^2$ can be written as:

$$\mathbf{v} = x \cdot \mathbf{b}_1 + y \cdot \mathbf{b}_2 = x \cdot [2, 0] + y \cdot [0, 3] = [2x, 3y]$$

  Therefore, $\{\mathbf{b}_1, \mathbf{b}_2\}$ forms a **basis** for $\mathbb{R}^2$.
- However, these vectors are **not unit vectors**, because:

$$\|\mathbf{b}_1\| = \sqrt{2^2 + 0^2} = 2 \quad \text{and} \quad \|\mathbf{b}_2\| = \sqrt{0^2 + 3^2} = 3$$

Thus, even though $\{\mathbf{b}_1, \mathbf{b}_2\}$ forms a valid basis, the vectors are **not unit vectors**.

Note: Two vectors b1 and b2 are **linearly independent** if the only solution to the equation c1 b1+c2 b2=0 is c1=0 and c2=0, where c1 and c2 are scalars.

Important points:

1. Basis Vectors Are Not Necessarily Unit Vectors

2. Unit Vectors Are Not Always a Basis

3. Unit Vectors Can Form a Basis if They Are Linearly Independent

Norms of Vectors for Data Science:

1. L1 Norm (Manhattan Normor Taxicab Norm):

The **L1 norm** measures the "absolute sum" of the components of a vector. It is called the **Manhattan norm** because it mimics the way a taxi would travel in a city grid (i.e., along the streets, not in a straight line).

Definition:

For a vector $\mathbf{v} = [v_1, v_2, \ldots, v_n]$, the L1 norm is:

$$\|\mathbf{v}\|_1 = |v_1| + |v_2| + \cdots + |v_n|$$

Example:

For $\mathbf{v} = [3, -4, 2]$,

$$\|\mathbf{v}\|_1 = |3| + |-4| + |2| = 3 + 4 + 2 = 9$$

2. L2 Norm (Euclidean Norm):

The **L2 norm** is the most commonly used norm in data science. It measures the straight-

Definition:

For a vector $\mathbf{v} = [v_1, v_2, \ldots, v_n]$, the **L2 norm** is:

$$\|\mathbf{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

Example:

For $\mathbf{v} = [3, 4]$,

$$\|\mathbf{v}\|_2 = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = \sqrt{25} = 5$$

line distance between a vector and the origin in the Euclidean space.

3. **L∞ Norm (Infinity Norm):** The L∞ norm, also known as the maximum norm or Chebyshev norm, measures the largest absolute value among the components of the vector. It focuses on the largest deviation in any direction.

Definition:

For a vector $\mathbf{v} = [v_1, v_2, \ldots, v_n]$, the **L∞ norm** is:

$$\|\mathbf{v}\|_\infty = \max(|v_1|, |v_2|, \ldots, |v_n|)$$

Example:

For $\mathbf{v} = [3, -4, 2]$,

$$\|\mathbf{v}\|_\infty = \max(|3|, |-4|, |2|) = 4$$

Important points:

1. The L1 norm encourages **sparse solutions**. In machine learning, this property is often used for **feature selection**, as it can push some coefficients (or weights) of a model to zero, effectively removing irrelevant features.

2. The L2 norm is often used because it provides a smooth, differentiable function that is useful for optimization algorithms (such as gradient descent).

3. The L2 norm gives more weight to larger values, making it sensitive to outliers in the data than L1.

4. The L∞ norm is useful when you want to **minimize the maximum deviation** in any direction.

Mathematical Example:

Consider two vectors:

- $\mathbf{v}_1 = [3, -4, 2]$
- $\mathbf{v}_2 = [1, 1, 1]$

The L∞ norms are:

$$\|\mathbf{v}_1\|_\infty = \max(|3|, |-4|, |2|) = 4$$

$$\|\mathbf{v}_2\|_\infty = \max(|1|, |1|, |1|) = 1$$

Here, $\mathbf{v}_1$ has a larger maximum deviation (4), so if we were to minimize the L∞ norm (the maximum deviation), we'd focus on reducing the largest component, which is 4 in this case. The L∞ norm directly addresses the "largest" component without regard to the smaller values in the vector.

Matrix

A **matrix** is a rectangular array of numbers or other mathematical objects, arranged in rows and columns. It can be used to represent samples with multiple attributes in a compact form. It can also be used to represent linear equations in a compact and simple fashion.

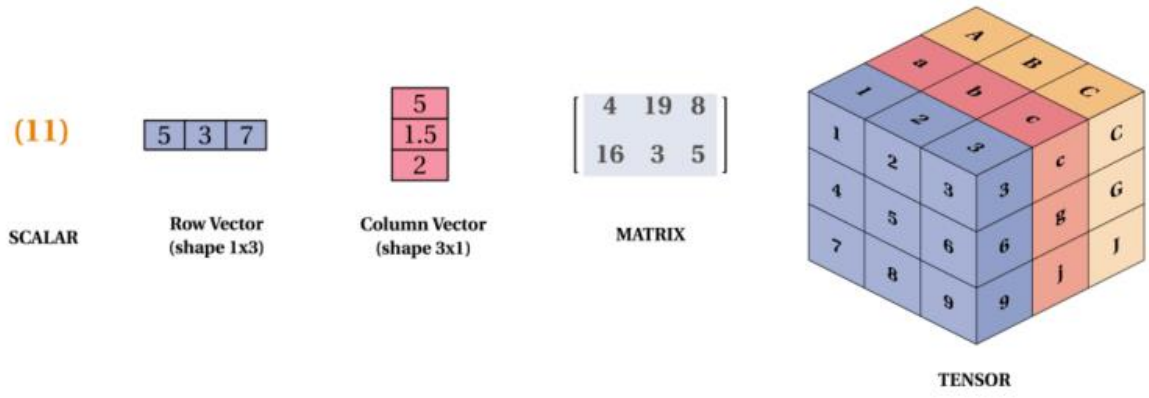
Definition and Notation:

A matrix is typically denoted by an uppercase letter, like A, and consists of elements arranged in rows and columns. The size of a matrix is expressed as m $\times$ n, where m is the number of rows and n is the number of columns.

For example, a matrix A with 2 rows and 3 columns looks like this:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Tensor: Generally, an n-dimensional array where n>2 is called a Tensor. But a matrix or a vector is also a valid tensor. A tensor is an algebraic object that describes a multilinear relationship between sets of algebraic objects related to a vector space.



Terms related to Matrix :

Order of matrix – If a matrix has 3 rows and 4 columns, order of the matrix is 3*4 i.e. row*column.

Square matrix – The matrix in which the number of rows is equal to the number of columns.

Diagonal matrix – A matrix with all the non-diagonal elements equal to 0 is called a diagonal matrix.

Upper triangular matrix – Square matrix with all the elements below diagonal equal to 0.

Lower triangular matrix – Square matrix with all the elements above the diagonal equal to 0.

Scalar matrix – Square matrix with all the diagonal elements equal to some constant k.

Identity matrix – Square matrix with all the diagonal elements equal to 1 and all the non-diagonal elements equal to 0.

Column matrix – The matrix which consists of only 1 column. Sometimes, it is used to represent a vector.

Row matrix – A matrix consisting only of row.

Trace – It is the sum of all the diagonal elements of a square matrix.

Baisc Operation on Matrix:

Addition – Addition of matrices is almost similar to basic arithmetic addition. Eg : Suppose we have 2 matrices 'A' and 'B' and the resultant matrix after the addition is 'C'. Then

$$Cij = Aij + Bij$$

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

Subtraction – Subtraction of matrices is almost similar to basic arithmetic subtraction. Eg : Suppose we have 2 matrices 'A' and 'B' and the resultant matrix after the subtraction is 'D'. Then

$$Dij = Aij - Bij$$

Let:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

The result of matrix subtraction $A - B$ is:

$$A - B = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \end{bmatrix}$$

Multiplication: Matrix multiplication is done by taking the dot product of rows of the first matrix with columns of the second matrix but the inner dimension should be same.

$$[A] \times [B] = [C]$$
$$(n \times m) \quad (m \times p) \quad (n \times p)$$

Inner dimensions need to be the same

The resulting matrix will be the outer dimensions

For the matrices A and B, we compute:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

The result of the matrix multiplication $A \times B$ is:

$$A \times B = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

Transpose of a Matrix:

The transpose of a matrix A, denoted as A^T, is formed by swapping the rows and columns of A.

Example:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \Rightarrow \quad A^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}$$

Email: sujan@ioepc.edu.np (for any Feedback)

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \rightarrow A^T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Properties:

$$(A^T)^T = A$$

$$(A + B)^T = A^T + B^T$$

$$(kA)^T = kA^T$$

$$(AB)^T = B^T A^T$$

Determinant of matrix:

The determinant of a matrix can be calculated from square matrices.

Evaluating Determinants

(1) Order One:

$$A = [\,a\,]$$

$$|A| = |\,a\,|$$

$$= a$$

(2) Order Two:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

(3) Order Three:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Properties of determinant:

1. The value of the determinant is unaltered by interchanging its rows and columns.
2. Interchanging any two adjacent rows (or columns) changes the sign of the determinant.
3. If any two rows (or columns) of a determinant are identical, then the value of the determinant is zero.
4. If all the elements of any row (or column) are multiplied by a constant k, then the value of the determinant is multiplied by k.
5. If each element of any row (or column) of a determinant is expressed as the sum of two terms, then the determinant can be expressed as a sum of two determinants.
6. If two the elements of any row (or column) a multiple of any other row (or column) is added, the value of the determinant remains unaltered.

Geometric Interpretation of the Determinant:

Magnitude of the Determinant:

The magnitude of the determinant tells us the size of the region spanned by the vectors.

For 2x2, this is the area of the parallelogram.

For 3x3, this is the volume of the parallelepiped(3-dimensional generalization of a parallelogram).

Sign of the Determinant:

The sign of the determinant indicates the orientation of the space.

Positive determinant: The space is oriented according to the right-hand rule (counterclockwise in 2D, right-handed in 3D).

Negative determinant: The space is reversed, following the left-hand rule (clockwise in 2D, left-handed in 3D).

Zero determinant: The vectors are linearly dependent, meaning they collapse to a lower dimension (no area or volume is formed).

Inverse of Matrix:

A square matrix A is invertible (or non-singular) if there exists another matrix A^{-1} such that $AA^{-1} = I$, where I is the identity matrix. Only square matrices can have an inverse.

Properties:

$$(A^{-1})^{-1} = A$$

$$A^{-1}A = AA^{-1} = I$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(A^{-1})^T = (A^T)^{-1}$$

Method 1: Using the determinant of the matrix

For 2*2 Matrix:

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{\det(A)}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

For 3*3 Matrix: Here C is Cofactor and M is Minors.

$$C_{ij} = (-1)^{i+j} M_{ij}$$

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \qquad \boldsymbol{M_{11}} = \det\left(\begin{bmatrix} e & f \\ g & h \end{bmatrix}\right), \qquad \boldsymbol{C_{11}} = (-1)^{1+1} M_{11} = M_{11}$$

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & -2 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\text{Firstly, find Cofactor}_A = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \text{ and } \det(A)$$

$$C_{11} = (-1)^{1+1}\left|\begin{bmatrix} 0 & -2 \\ 1 & 0 \end{bmatrix}\right| = 2, \quad C_{12} = (-1)^{1+2}\left|\begin{bmatrix} 0 & -2 \\ 0 & 0 \end{bmatrix}\right| = 0, \quad C_{13} = (-1)^{1+3}\left|\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right| = 0,$$

$$C_{21} = (-1)^{2+1}\left|\begin{bmatrix} 2 & 0 \\ 1 & 0 \end{bmatrix}\right| = 0, \quad C_{22} = (-1)^{2+2}\left|\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right| = 0, \quad C_{23} = (-1)^{2+3}\left|\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}\right| = -1,$$

$$C_{31} = (-1)^{3+1}\left|\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}\right| = -4, \quad C_{32} = (-1)^{3+2}\left|\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}\right| = 2, \quad C_{33} = (-1)^{3+3}\left|\begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}\right| = 0$$

$$\therefore \text{Cofactor}_A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & -1 \\ -4 & 2 & 0 \end{bmatrix}$$

$$\det(A) = a_{21}C_{21} + a_{22}C_{22} + a_{23}C_{23}$$

$$= 0 \times 0 + 0 \times 0 + -2 \times -1 = 2$$

$$A^{-1} = \frac{1}{\det(A)}\,\text{Cofactor}_A^T = \frac{1}{2}\begin{bmatrix} 2 & 0 & -4 \\ 0 & 0 & 2 \\ 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 0 & 1 \\ 0 & -0.5 & 0 \end{bmatrix}$$

Note: The determinant of a square matrix can be computed by multiplying entries in any row or columns by corresponding cofactors and adding the resulting products:

$$\det(A) = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in}, \qquad \text{for i} \le \text{i} \le \text{n (any row)}$$

or

$$\det(A) = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj}, \qquad \text{for i} \le \text{j} \le \text{n (any column)}$$

Method 2: Using elementary row operations

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 5 & 6 & 0 \end{bmatrix}$$

Step 1: Create the augmented matrix $[A|I]$:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 0 & 1 & 0 \\ 5 & 6 & 0 & 0 & 0 & 1 \end{array}\right]$$

$R_3 \to R_3 - 5R_1$:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 0 & 1 & 0 \\ 0 & -4 & -15 & -5 & 0 & 1 \end{array}\right]$$

Perform $R_3 \to R_3 + 4R_2$:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 4 & 0 & 1 & 0 \\ 0 & 0 & 1 & -5 & 4 & 1 \end{array}\right]$$

Perform $R_1 \to R_1 - 2R_2$:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -5 & 1 & -2 & 0 \\ 0 & 1 & 4 & 0 & 1 & 0 \\ 0 & 0 & 1 & -5 & 4 & 1 \end{array}\right]$$

$R_1 \to R_1 + 5R_3$:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -24 & 18 & 5 \\ 0 & 1 & 4 & 0 & 1 & 0 \\ 0 & 0 & 1 & -5 & 4 & 1 \end{array}\right]$$

$R_2 \to R_2 - 4R_3$:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -24 & 18 & 5 \\ 0 & 1 & 0 & 20 & -15 & -4 \\ 0 & 0 & 1 & -5 & 4 & 1 \end{array}\right]$$

$$A^{-1} = \begin{bmatrix} -24 & 18 & 5 \\ 20 & -15 & -4 \\ -5 & 4 & 1 \end{bmatrix}$$

Assignment: Find the inverse of the following matix using above method:

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & -2 \\ 0 & 1 & 0 \end{bmatrix}$$

Rank of a matrix :

Rank of a matrix is equal to the maximum number of linearly independent row vectors in a matrix. Rank is the number of rows with non zero vectors.The rank also indicates the number of non-zero eigenvalues in the matrix.

To Calculate Rank of Matrix:

1. If A is of order n × n and |A| not equal to 0, then the rank of A = n.

Given matrix D:

$$D = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 4 & 5 \\ 1 & 0 & 6 \end{pmatrix}$$

Determinant using cofactor expansion:

$$|D| = 2 \cdot \left| \begin{pmatrix} 4 & 5 \\ 0 & 6 \end{pmatrix} \right| - 1 \cdot \left| \begin{pmatrix} 0 & 5 \\ 1 & 6 \end{pmatrix} \right| + 3 \cdot \left| \begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix} \right|$$

After performing the necessary calculations for the 2x2 determinants (as shown previously), we get:

$$|D| = 48 + 5 - 12 = 41$$

Conclusion:

$$|D| = 41 \neq 0$$

This shows that matrix D has a non-zero determinant, confirming it is invertible and its rank is 3.

2. If C is of order n × n and |C| = 0, then the rank of C will be less than n.

$$C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Determinant of C:

$$|C| = 1 \cdot \left|\begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}\right| - 2 \cdot \left|\begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix}\right| + 3 \cdot \left|\begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix}\right|$$

Calculating the 2x2 determinants:

$$\left|\begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}\right| = (5 \times 9) - (6 \times 8) = 45 - 48 = -3$$

$$\left|\begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix}\right| = (4 \times 9) - (6 \times 7) = 36 - 42 = -6$$

$$\left|\begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix}\right| = (4 \times 8) - (5 \times 7) = 32 - 35 = -3$$

Substitute back into the determinant formula:

$$|C| = 1 \cdot (-3) - 2 \cdot (-6) + 3 \cdot (-3)$$

$$|C| = -3 + 12 - 9 = 0$$

Hence, the rank of C will be less than 3. Now take any 2*2 matrix form C and apply rule 1 and 2 again to find rank..Here for 2*2 matrix the |New 2*2 matirx| is not equal to 0, then the rank of C= 2.

3. If A matrix is of order m × n, then ρ(A) ⩽ min{m, n } = minimum of m, n

4. Rank of a Matrix by Row- Echelon Form :

A matrix is said to be in row-echelon form if the following rules are satisfied.

- All the leading entries in each row of the matrix is 1.
- If a column contains a leading entry then all the entries below the leading entry should be zero
- All rows which consist only of zeros should occur in the bottom of the matrix.
- the leading entry in the upper row should occur to the left of the leading entry in the lower row.

Example 1:

Find the rank of the matrix $A=\begin{pmatrix}1&1&1&1\\3&4&5&2\\2&3&4&0\end{pmatrix}$

Solution:

The order of A is 3×4.

$\therefore \rho(A) \leq 3$.

Let us transform the matrix A to an echelon form

Matrix A	Elementary Transformation
$A=\begin{pmatrix}1&1&1&1\\3&4&5&2\\2&3&4&0\end{pmatrix}$	
$\sim\begin{pmatrix}1&1&1&1\\0&1&2&-1\\0&1&2&-2\end{pmatrix}$	$R_2 \to R_2 - 3R_1$ $R_3 \to R_3 - 2R_1$
$\sim\begin{pmatrix}1&1&1&1\\0&-1&-2&1\\0&0&0&-1\end{pmatrix}$	$R_3 \to R_3 - R_2$

The number of non zero rows is 3.

$\therefore \rho(A) = 3$.

Example 2:

Find the rank of the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 5 & 7 \end{bmatrix}$

Solution:

Given $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 5 & 7 \end{bmatrix}$

Now we transform the matrix A to echelon form by using elementary transformation.

$R_2 \rightarrow R_2 - 2R_1$

$R_3 \rightarrow R_3 - 3R_1$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -2 \\ 0 & -1 & -2 \end{bmatrix}$$

$R_3 \rightarrow R_3 - R_2$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

Number of non-zero rows = 2

Hence the rank of matrix A = 2

Null Space and Nulity of a matrix:

The null space (or kernel) of a matrix A consists of all the vectors x that satisfy the equation: $Ax = 0$

In other words, the null space is the set of all vectors that, when multiplied by the matrix A, result in the zero vector.

The nullity of a matrix A is the dimension of its null space. It tells you the number of free variables (or degrees of freedom) in the solution to the homogeneous system Ax=0.

Rank-Nullity Theorem

The Rank-Nullity Theorem is an important result that connects the rank and nullity of a matrix. It states:

Rank(A)+Nullity(A)=n

Example:

Consider the following matrix A:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

We are tasked with finding the **null space** and the **nullity** of this matrix.

Step 1: Find the Null Space of A

We want to solve the homogeneous system:

$$A\mathbf{x} = \mathbf{0}$$

or

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

We will perform **row reduction** to find the solutions.

1. **Row 2** is just 2 times **Row 1**, so we can eliminate Row 2 by subtracting $2 \times \text{Row } 1$ from Row 2:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

2. Now, we subtract Row 1 from Row 3 to eliminate the first entry of Row 3:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & -2 & -3 \end{pmatrix}$$

3. Next, we multiply Row 3 by -1 to simplify:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}$$

Step 2: Solve for the Null Space

Now, we have the system:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This system gives the following equations:

$$x_1 + 2x_2 + 3x_3 + 4x_4 = 0$$
$$x_2 + 2x_3 + 3x_4 = 0$$

From the second equation, solve for x_2:

$$x_2 = -2x_3 - 3x_4$$

Substitute this into the first equation:

$$x_1 + 2(-2x_3 - 3x_4) + 3x_3 + 4x_4 = 0$$
$$x_1 - 4x_3 - 6x_4 + 3x_3 + 4x_4 = 0$$
$$x_1 = x_3 + 2x_4$$

Thus, the general solution is:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_3 + 2x_4 \\ -2x_3 - 3x_4 \\ x_3 \\ x_4 \end{pmatrix}$$

This can be rewritten as:

$$x_3 \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}$$

So the null space is spanned by the vectors:

$$\begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}$$

The nullity of A is the dimension of the null space. Since the null space is spanned by 2 vectors, the nullity of A is: Nullity(A)=2

Eigen Values and Eigen Vectors:

Eigenvalues represent scalar values that scale eigenvectors after matrix transformation

Eigenvectors are non-zero vectors scaled by eigenvalues when multiplied by a matrix

Given a square matrix A ($n \times n$),The mathematical formulation is, $Ax = \lambda x$
where, The constant λ (positive) represents the amount of stretch or shrinkage the attributes x go through in the x direction
The solution x are known as eigen vectors and λ is eigen values.

The equation $|A - \lambda I| = 0$ is called the characteristic equation of the matrix A

1. Eigenvectors represent directions in which the transformation (represented by A) acts by stretching or compressing. They remain in the same direction after the transformation.
Explaination: Imagine a matrix A that represents a transformation, such as a rotation, reflection, scaling, or some combination of these. When you multiply A by a vector x, you're applying this transformation to x, which generally results in a new vector in a different direction. However, eigenvectors are special vectors that, when multiplied by A, don't change direction.

2. Eigenvalues represent the scaling factors along these eigenvector directions. If $\lambda > 1$, the eigenvector is stretched. If $0 < \lambda < 1$, it is compressed. If $\lambda = -1$, the direction of the eigenvector is reversed.

Steps to find out the eigen value and eigen vector for variable x,

1. Find the characteristics equation $|A - \lambda I| = 0$

2. Solve the characteristics equation to get characteristic roots. They are called Eigen values.

3. To find the Eigen vectors, solve $|A - \lambda I|X = 0$ for different values of λ

For More Info:
https://kanchiuniv.ac.in/coursematerials/Eigenvalues_and_Eigenvectors.pdf

Example:

$$A = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix}$$

The eigenvalues of a matrix A are the solutions to the characteristic equation:

$$\det(A - \lambda I) = 0$$

where λ is the eigenvalue and I is the identity matrix.

First, compute $A - \lambda I$:

$$A - \lambda I = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1-\lambda & -1 & 4 \\ 3 & 2-\lambda & -1 \\ 2 & 1 & -1-\lambda \end{pmatrix}$$

Next, compute the determinant of this matrix:

$$\det(A - \lambda I) = \begin{vmatrix} 1-\lambda & -1 & 4 \\ 3 & 2-\lambda & -1 \\ 2 & 1 & -1-\lambda \end{vmatrix}$$

So, the characteristic polynomial is:

$$-\lambda^3 + 2\lambda^2 + 5\lambda - 6 = 0$$

Eigenvalue 1: λ=1

We substitute $\lambda = 1$ into the equation $(A - \lambda I)\mathbf{v} = 0$:

$$A - I = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 4 \\ 3 & 1 & -1 \\ 2 & 1 & -2 \end{pmatrix}$$

Now, solve the equation:

$$\begin{pmatrix} 0 & -1 & 4 \\ 3 & 1 & -1 \\ 2 & 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives the system of equations:

1. $-x_2 + 4x_3 = 0$
2. $3x_1 + x_2 - x_3 = 0$
3. $2x_1 + x_2 - 2x_3 = 0$

Solving by cross multiplication method:

$$\begin{array}{cccc} & x_1 & x_2 & x_3 \\ & -1 & 4 & -1 \\ & 1 & -1 & 1 \end{array}$$

$$\frac{x_1}{1-4} = \frac{x_2}{12-0} = \frac{x_3}{0+3}$$

$$\Rightarrow \frac{x_1}{-3} = \frac{x_2}{12} = \frac{x_3}{3}$$

$$\Rightarrow \frac{x_1}{-1} = \frac{x_2}{4} = \frac{x_3}{1}$$

$$\text{Therefore } X_1 = \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}$$

Eigenvalue 2: $\lambda = -2$

For $\lambda = -2$, we substitute into the equation $(A + 2I)\mathbf{v} = 0$:

$$A + 2I = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 3 & -1 & 4 \\ 3 & 4 & -1 \\ 2 & 1 & 1 \end{pmatrix}$$

Now, solve the equation:

$$\begin{pmatrix} 3 & -1 & 4 \\ 3 & 4 & -1 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives the system of equations:

1. $3x_1 - x_2 + 4x_3 = 0$
2. $3x_1 + 4x_2 - x_3 = 0$
3. $2x_1 + x_2 + x_3 = 0$

Solving by cross multiplication method:

$$\begin{array}{ccc} x_1 & x_2 & x_3 \\ -1 & 4 & 3 & -1 \\ 4 & -1 & 3 & 4 \end{array}$$

$$\frac{x_1}{1 - 16} = \frac{x_2}{12 + 3} = \frac{x_3}{12 + 3}$$

$$\Rightarrow \frac{x_1}{-15} = \frac{x_2}{15} = \frac{x_3}{15}$$

$$\Rightarrow \frac{x_1}{-1} = \frac{x_2}{1} = \frac{x_3}{1}$$

Therefore $X_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$

Eigenvalue 3: $\lambda = 3$

For $\lambda = 3$, we substitute into the equation $(A - 3I)\mathbf{v} = 0$:

$$A - 3I = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} - \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} -2 & -1 & 4 \\ 3 & -1 & -1 \\ 2 & 1 & -4 \end{pmatrix}$$

Now, solve the equation:

$$\begin{pmatrix} -2 & -1 & 4 \\ 3 & -1 & -1 \\ 2 & 1 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives the system of equations:

1. $-2x_1 - x_2 + 4x_3 = 0$
2. $3x_1 - x_2 - x_3 = 0$
3. $2x_1 + x_2 - 4x_3 = 0$

Solving by cross multiplication method:

$$\begin{array}{ccccc} & x_1 & & x_2 & x_3 \\ & -1 & 4 & -2 & -1 \\ & -1 & -1 & 3 & -1 \end{array}$$

$$\frac{x_1}{1+4} = \frac{x_2}{12-2} = \frac{x_3}{2+3}$$

$$\Rightarrow \frac{x_1}{5} = \frac{x_2}{10} = \frac{x_3}{5}$$

$$\Rightarrow \frac{x_1}{1} = \frac{x_2}{2} = \frac{x_3}{1}$$

Therefore $X_3 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$

Matrix factorization:

Matrix factorization is a mathematical technique used to decompose a matrix into a product of simpler matrices, each capturing distinct information about the structure of the original matrix. The purpose of matrix factorization is to reduce complex data structures into simpler components, which can be beneficial for tasks such as dimensionality reduction, data compression, and identifying latent factors within datasets.

Some types of Matrix factorization:
1. LU Decomposition
2. QR Decomposition
3. Cholesky Decomposition
4. SVD (Singular Value Decomposition)
5. Non-Negative Matrix Factorization (NMF)

1. LU Decomposition:

LU decomposition breaks down a square matrix into lower and upper triangular matrices. This technique simplifies solving linear equations and finding determinants, making it a key tool in linear algebra and data science applications.

Specifically, for a square matrix A, we write: $A=LU$

where, L is a lower triangular matrix with all elements on and below the diagonal, and the diagonal elements are typically set to 1.

And, U is an upper triangular matrix with all elements above the diagonal, and the diagonal elements can vary.

Note: Determinant of matrix A equals the product of U's diagonal elements.

Example:

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = LU \text{ where } L = \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}.$$

Multiplying out LU and setting the answer equal to A gives

$$\begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} & L_{21}U_{13} + U_{23} \\ L_{31}U_{11} & L_{31}U_{12} + L_{32}U_{22} & L_{31}U_{13} + L_{32}U_{23} + U_{33} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix}.$$

Now we use this to find the entries in L and U. Fortunately this is not nearly as hard as it might at first seem. We begin by running along the top row to see that

$$\boxed{U_{11} = 1}, \qquad \boxed{U_{12} = 2}, \qquad \boxed{U_{13} = 4}.$$

Now consider the second row

$$L_{21}U_{11} = 3 \quad \therefore L_{21} \times 1 = 3 \quad \therefore \boxed{L_{21} = 3},$$

$$L_{21}U_{12} + U_{22} = 8 \quad \therefore 3 \times 2 + U_{22} = 8 \quad \therefore \boxed{U_{22} = 2},$$

$$L_{21}U_{13} + U_{23} = 14 \quad \therefore 3 \times 4 + U_{23} = 14 \quad \therefore \boxed{U_{23} = 2}.$$

Notice how, at each step, the equation being considered has only one unknown in it, and other quantities that we have already found. This pattern continues on the last row

$$L_{31}U_{11} = 2 \quad \therefore L_{31} \times 1 = 2 \quad \therefore \boxed{L_{31} = 2},$$

$$L_{31}U_{12} + L_{32}U_{22} = 6 \quad \therefore 2 \times 2 + L_{32} \times 2 = 6 \quad \therefore \boxed{L_{32} = 1},$$

$$L_{31}U_{13} + L_{32}U_{23} + U_{33} = 13 \quad \therefore (2 \times 4) + (1 \times 2) + U_{33} = 13 \quad \therefore \boxed{U_{33} = 3}.$$

We have shown that

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

and this is an LU decomposition of A.

2. QR Decomposition:

QR decomposition is a method of decomposing a matrix A into the product of two matrices Q and R,
where, Q is an orthogonal matrix (i.e., $Q^TQ=I$, where I is the identity matrix), and R is an upper triangular matrix.

The Gram-Schmidt process is one of the most common ways to perform QR decomposition. It involves orthogonalizing the columns of A to form the matrix Q and then finding the matrix R.

Gram-Schmidt Process:

Let A = [a 1 , a 2 , ⋯ , a n] be the matrix with columns a1 , a2 , ⋯ , an. The Gram-Schmidt process produces the orthogonal vectors [q 1 , q 2 , ⋯ , q n] and the upper triangular matrix R .

1. **Step 1: Find q_1:** Normalize the first column a_1 to form the first vector of Q:

$$q_1 = \frac{a_1}{\|a_1\|}$$

2. **Step 2: Find q_2:** Subtract the projection of a_2 onto q_1 to make it orthogonal to q_1:

$$u_2 = a_2 - \text{proj}_{q_1}(a_2) = a_2 - (\langle a_2, q_1 \rangle) q_1$$

Normalize u_2 to form q_2:

$$q_2 = \frac{u_2}{\|u_2\|}$$

3. **Step 3: Repeat for the remaining columns:** For each subsequent column a_k, subtract the projections of a_k onto all the previously computed orthogonal vectors:

$$u_k = a_k - \sum_{i=1}^{k-1} \langle a_k, q_i \rangle q_i$$

Then normalize u_k to form q_k:

$$q_k = \frac{u_k}{\|u_k\|}$$

4. **Step 4: Construct the matrix Q:** The columns of Q are the orthonormal vectors $q_1, q_2, \ldots, q_n$.
5. **Step 5: Construct the matrix R:** The matrix R is an upper triangular matrix where the i-th column of R is given by:

$$r_{ij} = \langle a_j, q_i \rangle \quad \text{for} \quad 1 \leq i \leq j \leq n$$

Example 1:

Example:

Let's do a simple QR decomposition on the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

1. **Matrix A:**

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

2. **Step 1: Compute q_1:** Normalize the first column of A:

$$q_1 = \frac{1}{\sqrt{1^2 + 3^2}} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

So,

$$q_1 = \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{pmatrix}$$

3. **Step 2: Compute q_2:** Subtract the projection of the second column $a_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ onto q_1:

$$\text{proj}_{q_1}(a_2) = \langle a_2, q_1 \rangle q_1 = \left(\frac{2}{\sqrt{10}} + \frac{12}{\sqrt{10}} \right) \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{pmatrix} = \frac{14}{10} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

This simplifies to:

$$\text{proj}_{q_1}(a_2) = \begin{pmatrix} 1.4 \\ 4.2 \end{pmatrix}$$

Now subtract this projection from a_2:

$$u_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 1.4 \\ 4.2 \end{pmatrix} = \begin{pmatrix} 0.6 \\ -0.2 \end{pmatrix}$$

Normalize u_2:

$$q_2 = \frac{1}{\sqrt{0.6^2 + (-0.2)^2}} \begin{pmatrix} 0.6 \\ -0.2 \end{pmatrix} = \frac{1}{\sqrt{0.36 + 0.04}} \begin{pmatrix} 0.6 \\ -0.2 \end{pmatrix} = \frac{1}{\sqrt{0.4}} \begin{pmatrix} 0.6 \\ -0.2 \end{pmatrix}$$

So:

$$q_2 = \begin{pmatrix} \frac{3}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \end{pmatrix}$$

4. **Step 3: Construct Q and R:**

   The orthogonal matrix Q is:

$$Q = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{pmatrix}$$

   The upper triangular matrix R is:

$$R = \begin{pmatrix} \sqrt{10} & \frac{14}{\sqrt{10}} \\ 0 & \sqrt{0.4} \end{pmatrix}$$

Thus, the QR decomposition of A is:

$$A = QR = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{pmatrix} \begin{pmatrix} \sqrt{10} & \frac{14}{\sqrt{10}} \\ 0 & \sqrt{0.4} \end{pmatrix}$$

Here, the R is upper triangular and contains the dot products of the columns of A with q1, q2, and q3.

For more info:
https://www.slideshare.net/slideshow/matrixdecompositionanditsapplicationinstatisticsnkppt/267316462

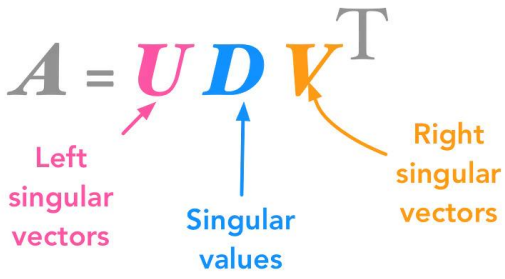
https://docs.google.com/presentation/d/1-aF2ONsjkTJjiq4qetmfMywhdnhcoiqr/edit#slide=id.p107

4. SVD (Singular Value Decomposition)

Singular value decomposition (SVD) is a matrix factorization method that generalizes the eigendecomposition of a square matrix (n x n) to any matrix (n x m). It factors a matrix into one diagonal matrix and two orthogonal matrices.

SVD is similar to principal component analysis (PCA), but it's more general. PCA assumes that the input is a square matrix, SVD doesn't have this assumption.

The general formula for SVD is:



where:

A: Original matrix (size m × n).

M is the original matrix we want to decompose.

U is the left singular matrix (columns are left singular vectors). U columns contain eigenvectors of matrix MM^t. (size m × m)

Σ is a diagonal matrix containing singular eigenvalues. (size m × n) Singular values are singular values are the square roots of the eigenvalues of AA^T. Also, non-negative and sorted in descending order.

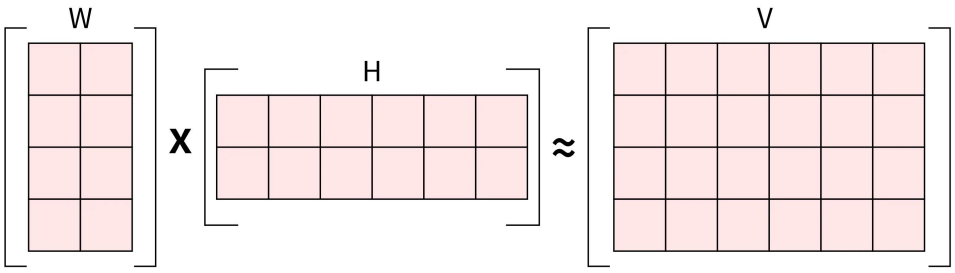
V is a right singular matrix (columns are right singular vectors). V columns contain eigenvectors of matrix M^tM. (size n × n).

Note: U and V are orthogonal matrices: $U^TU=I$ and $V^TV=I$, where I is the identity matrix.

5. Non-Negative Matrix Factorization (NMF)

Researchpaper:
https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf

Non-Negative Matrix Factorization (NMF) is a technique in linear algebra used for dimensionality reduction, clustering, and feature extraction. The key feature of NMF is that it decomposes a given non-negative matrix V into two non-negative matrices W and H, such that:



Where:

$V \in R^{m \times n}$ is the input non-negative matrix (e.g., data matrix with mm features and nn samples).

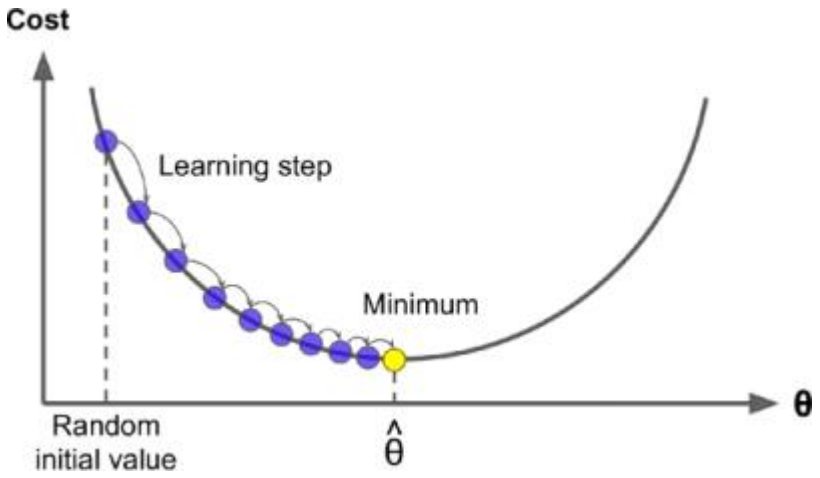
$W \in R^{m \times k}$ is the basis matrix (contains kk latent features or components).

$H \in R^{k \times n}$ is the encoding matrix (represents the coefficients or weights for each component).

2.3 Gradient Descent for Optimization:

Gradient Descent is a generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of Gradient Descent is to tweak parameters iteratively in order to minimize a cost function.

Suppose we have to minimize the value of "W" for a given loss function. Concretely, we start by filling "W" with random values (random initialization). Then we improve it gradually, taking one baby step at a time, each step attempting to decrease the loss function, until the algorithm converges to a minimum.



Let's first try to grab the concept with a raw example, keeping all ML algorithms aside.

Let's say we have an equation $y = 3x^2 - 2x + 5$. By solving this equation,

$$\frac{dy}{dx} = 6x - 2$$

$$\frac{dy}{dx} = \begin{cases} +_{ve}, & \text{when } x > \frac{1}{3} \\ 0, & \text{when } x = \frac{1}{3} \\ -_{ve}, & \text{when } x < \frac{1}{3} \end{cases}$$

and we know that minimum values exist when $\frac{dy}{dx} = 0$, so y will be minimum at $x = \frac{1}{3}$

Update rule for Gradient Descent:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla f(\theta)$$

Where:

- θ_{new} is the updated value of the parameter vector after applying the gradient descent step.
- θ_{old} is the current value of the parameter vector.
- η is the **learning rate**, a scalar value that determines how large a step we take in the direction of the negative gradient.
- $\nabla f(\theta)$ is the **gradient** of the cost function $f(\theta)$ with respect to the parameters θ, which gives the direction of steepest ascent (so we subtract it to go in the direction of steepest descent).

Minimizing a Simple Quadratic Function with Gradient Descent

Let's consider a quadratic function:

$f(x)=ax^2+bx+c$, $a>0$

Gradient descent helps us find the value of xx that minimizes f(x).

Steps for Minimizing f(x):

1. **Compute the Gradient:** The gradient of f(x) is the derivative of f(x) with respect to x:

$$\nabla f(x) = \frac{d}{dx}[ax^2 + bx + c] = 2ax + b$$

2. **Gradient Descent Update Rule:** To minimize f(x), we iteratively update xx using the formula:

$$x_{t+1} = x_t - \eta \cdot \nabla f(x_t)$$

Here:

- x_t is the value of x at iteration t,
- η is the learning rate (step size),
- $\nabla f(x_t) = 2ax_t + b$.

3. Iterative Steps:

- Initialize x_0 (starting point).
- Compute the gradient $\nabla f(x_t)=2ax_t+b$
- Update x using the gradient descent rule: $x_{t+1}=x_t- \eta\ (2ax_t+b)$
- Repeat until the change in xx (or the gradient) becomes very small.

Note: Gradient descent converges to the minimum of f(x) if the learning rate η is chosen appropriately. If η is too large, it may oscillate or diverge. If it's too small, convergence will be slow.

Example

Let's minimize $f(x) = x^2 - 4x + 6$:

1. **Function:** $f(x) = x^2 - 4x + 6$
2. **Gradient:** $\nabla f(x) = 2x - 4$
3. **Update Rule:**

$$x_{t+1} = x_t - \eta(2x_t - 4)$$

Iterations:

1. Initialize $x_0 = 0$, learning rate $\eta = 0.1$.

2. Iteration 1:

$$\nabla f(x_0) = 2(0) - 4 = -4$$

Update:

$$x_1 = 0 - 0.1(-4) = 0.4$$

3. Iteration 2:

$$\nabla f(x_1) = 2(0.4) - 4 = -3.2$$

Update:

$$x_2 = 0.4 - 0.1(-3.2) = 0.72$$

4. Continue iterating until convergence (gradient close to zero).

Minimizing Mean Square Error (MSE) with Gradient Descent

Mean Square Error (MSE) is commonly used as a loss function in regression problems. It measures the average squared difference between the predicted values and the actual values in a dataset.

For a set of predictions $\hat{y}_i$ and true labels y_i, the MSE is given as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Where:

- n is the number of data points.
- $\hat{y}_i = w \cdot x_i + b$ is the predicted value for the i-th data point, with model parameters w (weight) and b (bias).
- x_i is the input for the i-th data point.

Steps for minimizing MSE with Gradient Descent:

1. Initialize Parameters: Start with random values for w and b.

2. Compute the Gradient: For each parameter, compute the partial derivative of the MSE cost function.

- Partial derivative w.r.t. w:

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - \hat{y}_i)$$

- Partial derivative w.r.t. b:

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)$$

3. Update Parameters: Update w and b using the gradient and a learning rate η:

$$w = w - \eta \cdot \frac{\partial \text{MSE}}{\partial w}$$
$$b = b - \eta \cdot \frac{\partial \text{MSE}}{\partial b}$$

4. Iterate: Repeat steps 2 and 3 until it return the optimized w and b.

Example:

We use the following data points:

x	y
1	2
2	4
3	6

We'll use **gradient descent** to estimate the parameters w (slope) and b (intercept).

Initialization:

w=0 (initial slope), b=0 (initial intercept), Learning rate η=0.1, Number of data points n=3.

Iteration 1:

1. **Predicted values:**

$$\hat{y}_i = w \cdot x_i + b = 0 \cdot x_i + 0 = 0$$

So, $\hat{y} = [0, 0, 0]$.

2. **Gradients:**

- For w:

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{n} \sum_{i=1}^{n} x_i \left(y_i - (w \cdot x_i + b)\right)$$

Substituting $w = 0, b = 0$:

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{3} [1(2-0) + 2(4-0) + 3(6-0)]$$

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{3}(2 + 8 + 18) = -\frac{2}{3}(28) = -18.67$$

- For b:

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{n} \sum_{i=1}^{n} \left(y_i - (w \cdot x_i + b)\right)$$

Substituting $w = 0, b = 0$:

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{3} [(2-0) + (4-0) + (6-0)]$$

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{3}(12) = -8$$

3. **Update Parameters:**

$$w \leftarrow w - \eta \cdot \frac{\partial \text{MSE}}{\partial w} = 0 - 0.1 \cdot (-18.67) = 1.867$$

$$b \leftarrow b - \eta \cdot \frac{\partial \text{MSE}}{\partial b} = 0 - 0.1 \cdot (-8) = 0.8$$

Updated Parameters: $w = 1.867, b = 0.8$

Iteration 2:

1. **Predicted values:**

$$\hat{y}_i = w \cdot x_i + b = 1.867 \cdot x_i + 0.8$$

Substituting $x = [1, 2, 3]$:

$$\hat{y} = [2.667, 4.534, 6.401]$$

2. **Gradients:**

- For w:

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{3}\left[1(2 - 2.667) + 2(4 - 4.534) + 3(6 - 6.401)\right]$$

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{3}\left[-0.667 - 1.068 - 1.203\right]$$

$$\frac{\partial \text{MSE}}{\partial w} = -\frac{2}{3}(-2.938) = 1.959$$

- For b:

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{3}\left[(2 - 2.667) + (4 - 4.534) + (6 - 6.401)\right]$$

$$\frac{\partial \text{MSE}}{\partial b} = -\frac{2}{3}(-1.938) = 1.292$$

3. **Update Parameters:**

$$w \leftarrow w - \eta \cdot \frac{\partial \text{MSE}}{\partial w} = 1.867 - 0.1 \cdot (1.959) = 1.671$$

$$b \leftarrow b - \eta \cdot \frac{\partial \text{MSE}}{\partial b} = 0.8 - 0.1 \cdot (1.292) = 0.671$$

Updated Parameters: $w = 1.671, b = 0.671$

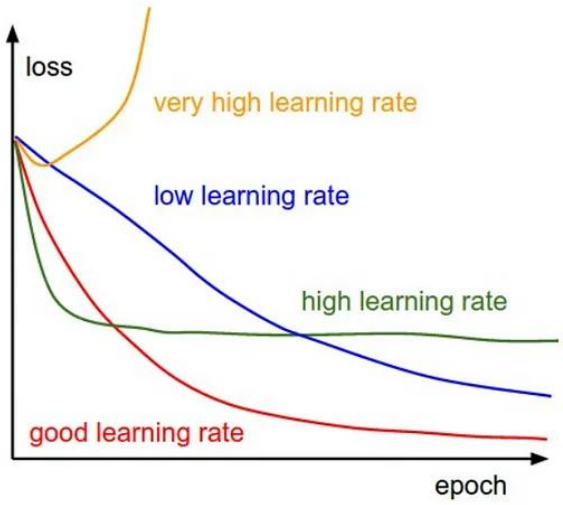
Continuing Iterations

Repeat this process for several iterations. After sufficient updates, the parameters converge to: $w \approx 2, b \approx 0$. This aligns with the true relationship $y = 2x$.

What we observed?

· Gradient descent reduces the error step by step.
· The learning rate (η) controls the speed of convergence.
· The process can be automated for many iterations until convergence or until the gradients become negligible.

Choosing Right Learning rate:



Gradient descent variants

1. Batch gradient descent(Normal Gradient Descent as mentioned above)

2. Stochastic gradient descent

3. Mini-batch gradient descent

1. Batch gradient descent: Vanilla gradient descent, aka batch gradient descent, computes the gradient of the cost function w.r.t. to the parameters θ for the entire training dataset:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta).$$

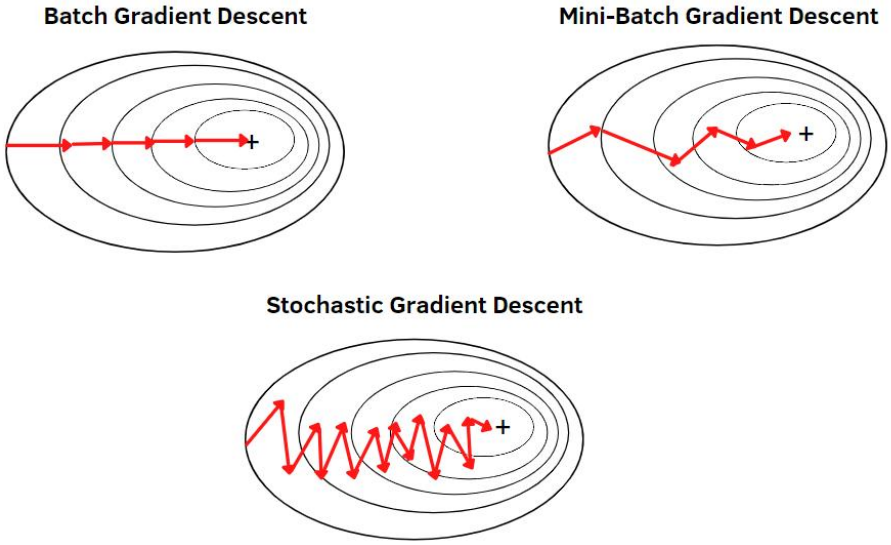
As we need to calculate the gradients for the whole dataset to perform just oneupdate, batch gradient descent can be very slow and is intractable for datasets that don't fit in memory.

2. Stochastic gradient descent: The main problem in Batch Gradient Descent is the fact that it still uses the whole training dataset to compute the gradients at every step, which makes it very slow when the training set is large.

Stochastic gradient descent (SGD) in contrast performs a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)}).$$

It is therefore usually much faster.



SGD performs frequent updates with a high variance that cause the objective function to fluctuate heavily. It is observed that in SGD the updates take more number iterations and is noiser compared to gradient descent to reach minima. On the right, the Gradient Descent takes fewer steps to reach minima.

3. Mini-Batch Gradient Descent:

Mini-batch gradient descent finally takes the best of both worlds and performs an update for every mini-batch of n training examples:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i:i+n)}; y^{(i:i+n)}).$$

This way, it a) reduces the variance of the parameter updates, which can lead to more stable convergence; and b) can make use of highly optimized matrix optimizations common to state-of-the-art deep learning libraries that make computing the gradient w.r.t. a mini-batch very efficient.

Gradient Descent optimization algorithms:

1. SGD with momentum

2. Nesterov Accelerated Gradient (NAG)

3. Adaptive Gradient (AdaGrad)

4. AdaDelta

5. RMSprop

6. Adam

Gradient descent challenges (local minima, overfitting, etc.)

1. Choosing the Learning Rate: Setting a learning rate that's too high can cause the algorithm to diverge, while too low of a rate may result in slow or suboptimal convergence.

2. Local Minima/Non-convex Loss Functions: In complex models, especially deep neural networks, the loss function is often non-convex, meaning it has many valleys and peaks. This can lead to difficulty in finding the global minimum, as the gradient descent might converge to different local minima depending on the initialization

3. Vanishing and Exploding Gradients
Vanishing Gradients: In deep neural networks, especially in those with many layers, the gradients of the loss function with respect to the model's parameters can become very small as they are propagated backward. This makes it difficult for the network to learn and update the parameters in the earlier layers. This is known as the vanishing gradient problem.
Exploding Gradients: Conversely, in some cases, the gradients can grow exponentially, leading to unstable updates and causing the model weights to become too large.

4. Computational Efficiency: For large datasets, computing gradients can be time-consuming.

5. Plateaus: Gradients near flat regions can be too small, slowing down progress.

Assignment: This is the loss function: $f(x,y)=(x-3)^2+(y+2)^2$; use the gradient descent optimization up to 5 iterations.

2.4 Introduction to probability and random variable:

- ✓ Outcome: A single result of an experiment (e.g., rolling a 4 on a die).
- ✓ Sample Space (S): The set of all possible outcomes of an experiment (e.g., for a die, S={1,2,3,4,5,6}.
- ✓ Event: A subset of the sample space. For example:
  - Event A: Rolling an even number (A={2,4,6}A={2,4,6}).
  - Event B: Rolling a number greater than 3 (B={4,5,6}B={4,5,6}).

Probability is a measure of the likelihood that an event will occur. For any event A, the probability of A, denoted as P(A), is defined as a number between 0 and 1, where:

P(A)=0 means A will never occur.

P(A)=1 means A is certain to occur.

Basic Rules of Probability

1. Additive Rule:

For two events A and B, the probability that A or B (or both) will occur is given by:

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

If A and B are mutually exclusive (cannot happen simultaneously), then $P(A \cap B)=0$, and the rule simplifies to: $P(A \cup B)=P(A)+P(B)$

2. Multiplicative Rule:

For two events A and B, the probability that both A and B will occur (joint probability) is: $P(A \cap B)=P(A) \cdot P(B \mid A)$

If A and B are independent events (the occurrence of A does not affect B and vice versa), this simplifies to:

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

Conditional Probability

The probability of an event A given that another event B has occurred is called the conditional probability, denoted P(A | B). It's defined by:

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

where $P(B) \neq 0$.

Imagine you conducted a survey where you asked 100 men and women of all ages if they eat meat and obtain the following results:

	Vegetarian	Not Vegetarian	Total
Women	15	32	47
Men	29	24	53
Total	44	56	100

Now, if A represents being vegetarian and B represents being a woman then,

$P(A|B) = \frac{15}{47}$ and $P(B|A) = \frac{15}{44}$

Baye's Theorem:

Bayes' theorem is a mathematical rule for inverting conditional probabilities, allowing us to find the probability of a cause given its effect.

Bayes' theorem describes the relationship between P(A | B) and P(B | A):

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

Numerical:

Suppose a medical test is used to detect a certain disease. 1% of the population has the disease. If a person has the disease, the test will correctly identify it 95% of the time.If a person does not have the disease, the test will incorrectly identify them as positive 5% of the time. Given that a person tests positive, what is the probability that they actually have the disease?

Given:

P(Disease)=0.01
P(Positive Test | Disease)=0.95
P(Positive Test | No Disease)=0.05

P(Disease | Positive Test)=?

By Bayes theorm:

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Disease}) \cdot P(\text{Disease})}{P(\text{Positive Test})}$$

Using the Law of Total Probability:

$$P(\text{Positive Test}) = P(\text{Positive Test} \mid \text{Disease}) \cdot P(\text{Disease}) + P(\text{Positive Test} \mid \text{No Disease}) \cdot P(\text{No Disease})$$

$$P(\text{Positive Test}) = (0.95 \cdot 0.01) + (0.05 \cdot 0.99)$$

$$P(\text{Positive Test}) = 0.0095 + 0.0495 = 0.059$$

Now, substitute all values into Bayes' Theorem:

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{0.95 \cdot 0.01}{0.059}$$

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{0.0095}{0.059} \approx 0.161$$

Hence, The probability that a person has the disease, given that they tested positive is 16.1%.

Assignment: A company runs an online advertising campaign and wants to know the probability that a customer who clicked on an ad actually makes a purchase. 10% of the customers who see the ad make a purchase. 80% of the people who make a purchase clicked on the ad. 5% of people who did not make a purchase clicked on the ad. What is the probability that a customer actually made a purchase, given that they clicked on the ad? (**Ans: 0.64 or 64%)**

Assignment: Use Bayes theorem to evaluate the probability of being sick given the diagnosis is positive for the following scenario: Out of 1,000,000 in a city, sick probability is 1/10,000; efficiency of the diagnosis tool is 99%

Hint: $P(+|S)=99\%=0.99$ and $P(+|\sim S)=1\%=0.01$

Random Variables:

A random variable is a variable whose possible values are numerical outcomes of a random phenomenon.

Types of Random Variable:

1. Discrete Random Variable: X is a discrete because it has a countable values between two numbers

Example : number of balls in a bag, number of tails in tossing coin

2. Continuous Random Variable: X is a continuous because it has a infinite number of values between two values

Example : distance travelled, Height of students

Expectation (Mean)

The expectation of a random variable X is the average value it takes in the long run, often denoted as E(X).

- For a **discrete random variable**:

$$E(X) = \sum_x x \cdot P(X = x)$$

- For a **continuous random variable**:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Variance and Standard Deviation

Variance measures the spread of a random variable around its mean. And The standard deviation is the square root of variance:

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right] \qquad \sigma_X = \sqrt{\mathrm{Var}(X)}$$

Probability Mass Function (PMF): The Probability Mass Function (PMF) is a function used to describe the probability distribution of a discrete random variable. It gives the probability that a discrete random variable X takes on a particular value x.

Definition:

For a discrete random variable X, the PMF is a function $p(x)=P(X=x)$ that satisfies the following properties:

1. Non-negativity: $p(x) \geqslant 0$ for all values of x. i.e The probability of any value x must be non-negative.

2. The sum of the probabilities for all possible values of X must equal 1.

PMF Example:

Let's consider the example of rolling a fair six-sided die. Let X be the outcome of the roll. Since the die has six faces, the PMF of X is:

$$p(x) = P(X = x) = \frac{1}{6}, \quad x \in \{1, 2, 3, 4, 5, 6\}$$

This means the probability of each value x (1 through 6) is $\frac{1}{6}$, and the sum of all probabilities is 1:

$$\sum_{x=1}^{6} p(x) = \sum_{x=1}^{6} \frac{1}{6} = 1$$

Probability Density Function (PDF):

For a continuous random variable X, the Probability Density Function (PDF), denoted as f(x), has the following properties:

1. Non-negativity: The PDF must be non-negative for all values of x.

2. Normalization: The total area under the PDF curve is 1.

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

3. Probability Calculation: Since the PDF represents a density, the probability that X lies within a specific range [a,b] is given by the area under the PDF curve between a and b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Note that for any exact value x, $P(X = x) = 0$ because the probability at any single point in a continuous distribution is infinitesimally small.

Cumulative Distribution Function (CDF): The Cumulative Distribution Function (CDF) of a random variable X provides the probability that X will take a value less than or equal to a particular value x. In other words, the CDF gives the cumulative probability up to a certain point.

CDF for Discrete Random Variables:

For a discrete random variable X with possible outcomes $x_1, x_2, \ldots$, the CDF is the cumulative sum of the probabilities up to x:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

Where $p(x_i)$ is the probability mass function (PMF) for discrete X.

CDF for Continuous Random Variables:

For continuous random variables, the CDF is the integral of the probability density function (PDF) up to x. The relationship is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Where $f(x)$ is the probability density function (PDF) of X.

2.5 Probability distributions: Normal, Bernoulli, Binomial, Poisson :

The probability distribution assigns a probability to each possible outcome of a random experiment, and it can be represented either by a probability mass function (PMF) for discrete variables or a probability density function (PDF)for continuous variables.

1. Bernoulli distributions: The Bernoulli Distribution is a discrete probability distribution representing a random experiment with exactly two outcomes: success (1) or failure (0). It's often used to model binary data or events.
Parameter : p is Probability of success ($0 \leqslant p \leqslant 1$).

The probability mass function (PMF) is:

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

Where:

- X is the random variable.
- p is the probability of success.

2. Binomial distributions:
The Binomial Distribution tells you the probability of having exactly k successes (e.g., heads) in n independent trials (e.g., coin flips), where each trial has a success probability of p.

Example: Coin Flips

Let's use the **coin flip** example:

- You flip a fair coin (which has a 50% chance of landing heads, so $p = 0.5$).
- You flip the coin **3 times**, so $n = 3$.
- You want to find the probability of getting exactly **2 heads** (successes), so $k = 2$.

In this case:

- The total number of trials is 3 (you flip the coin 3 times).
- The probability of getting heads (success) on each flip is $p = 0.5$.
- You want the probability of getting exactly 2 heads out of 3 flips.

How to Calculate the Probability

You would use the **Binomial formula**:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where:

- $P(X = k)$ is the probability of getting exactly k successes.
- $\binom{n}{k}$ is the **binomial coefficient** (which tells you the number of ways to choose k successes from n trials).
- p^k is the probability of getting k successes.
- $(1-p)^{n-k}$ is the probability of getting $n-k$ failures.

3. Poisson Distribution: The Poisson Distribution is a discrete probability distribution that models the number of events (or occurrences) happening in a fixed interval of time, space, or any other context, when the events occur independently and at a constant average rate. It is especially useful for modeling rare events that happen over a continuous domain, like the number of phone calls at a call center in an hour, the number of accidents at a particular intersection, or the number of emails received in a day.

The probability of exactly k events occurring in a fixed interval of time or space is given by the Poisson probability mass function:

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

Where:

- $P(X = k)$ is the probability of observing exactly k events.
- λ is the **average rate of events** in the interval (mean of the distribution).
- k is the number of events you are interested in (can be 0, 1, 2, ...).
- e is Euler's number (approximately 2.71828), which is the base of the natural logarithm.

Example:

Suppose a call center receives, on average, 3 calls per minute. This implies that the average rate λ of calls per minute is 3.

- If we want to know the probability of receiving exactly 5 calls in a minute, we would use the Poisson distribution formula with $\lambda = 3$ and $k = 5$.

Step-by-Step Calculation:

Using the Poisson formula:

$$P(X = 5) = \frac{3^5 e^{-3}}{5!}$$

1. **Calculate powers and factorial:**

$$3^5 = 243, \quad e^{-3} \approx 0.0498, \quad 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

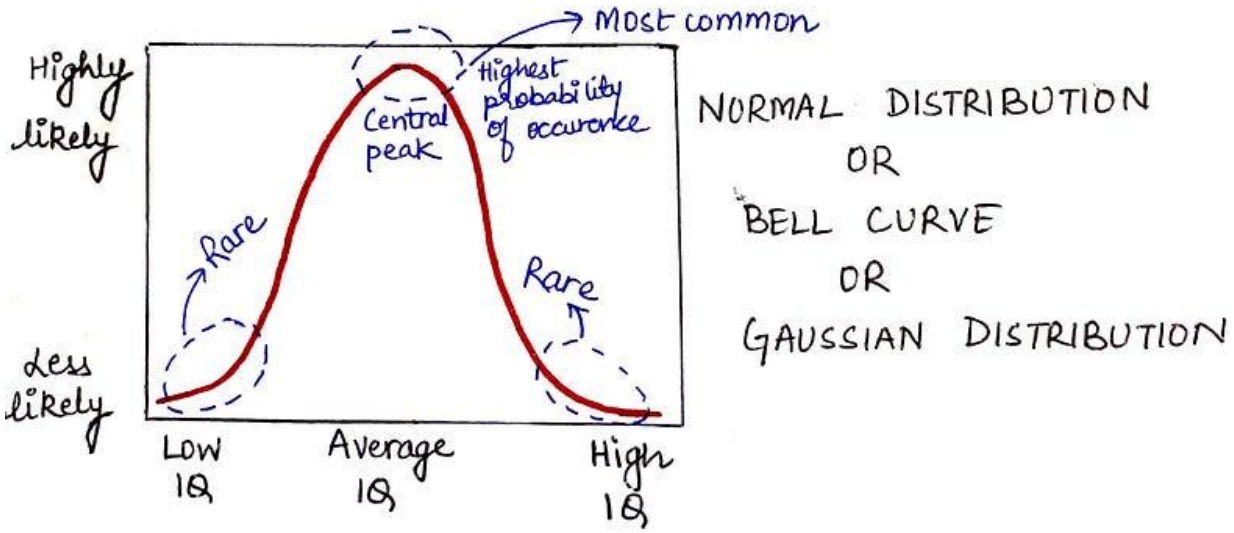
2. **Substitute into the formula:**

$$P(X = 5) = \frac{243 \times 0.0498}{120}$$

$$P(X = 5) \approx \frac{12.1}{120} \approx 0.1008$$

So, the probability of receiving exactly 5 calls in a minute is approximately **0.1008** (or 10.08%).

4. Normal Distrubution: The normal distribution is also known as Gaussian distribution or bell curve. For example, heights, weights, blood pressure, measurement errors, IQ scores etc. all follow the normal distribution.

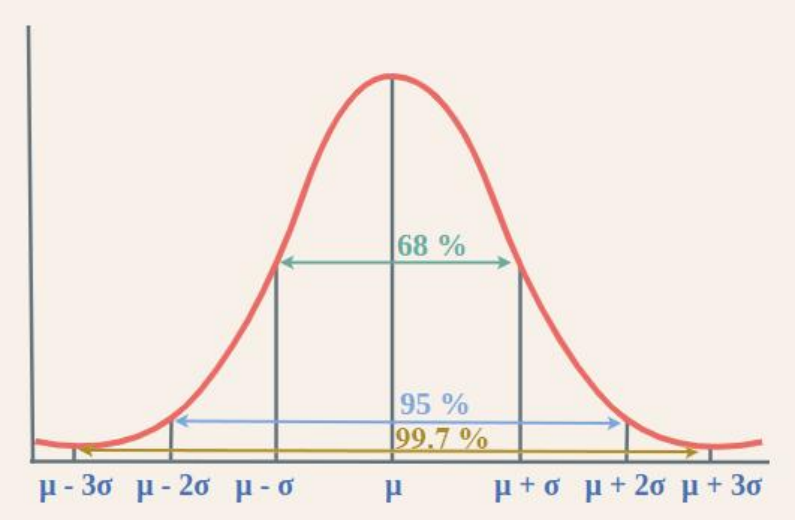


In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

It is Defined by Two Parameters:

Mean (μ): The average or central value of the distribution.

Standard Deviation (σ): Measures the spread of the distribution. A larger standard deviation means the data points are more spread out, while a smaller standard deviation means they are closer to the mean.



Studying the graph it is clear that using Empirical Rule we distribute data broadly in three parts. And thus, empirical rule is also called "68 – 95 – 99.7" rule.

The formula for the probability density function of the normal distribution is:

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

Where:

- $f(x)$ is the value of the probability density function at point x.
- μ is the **mean** of the distribution.
- σ is the **standard deviation** of the distribution.
- e is Euler's number (approximately 2.71828).

Standard Normal Distribution: A Standard Normal Distribution is a normal distribution with Mean μ=0 and Standard Deviation σ=1. The random variable of a

standard normal distribution is known as the standard score or a z-score. It is possible to transform every normal random variable X into a z score using the following formula:

$z = (X - \mu) / \sigma$

where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X.

Use cases of Standard Normal Distribution:

1. If you want to compare two exam scores from different tests, each with different means and standard deviations, converting them to Z-scores allows for a meaningful comparison, because they are now on the same scale.

2. Calculating cumulative probability, Hypothesis testing, Comparing Different Data Sets, Quality control and process monitoring in manufacturing, Assessing risk and uncertainty in fields like finance etc.

1. The weights of bags of flour in a factory follow a normal distribution with a mean of 4.76 kg and a standard deviation of 0.05 kg.

- ✓ What is the probability that a randomly selected bag of flour weighs between 4.6 kg and 4.8 kg?
- ✓ What is the probability that a randomly selected bag weighs more than 4.8 kg?
- ✓ What is the probability that a randomly selected bag weighs less than 4.6 kg?

SOLUTION:

Mean μ=4.76 kg and Standard deviation σ=0.05 kg

Use the Z-score formula to convert the values to standard normal variables.

1. probability that a randomly selected bag of flour weighs between 4.6 kg and 4.8 kg = P(4.6<X<4.8)

using z score formula,

= P(4.6<X<4.8)

$= P(\frac{4.6-4.76}{0.05} < \frac{X-\mu}{\sigma} < \frac{4.8-4.76}{0.05})$

= P(-3.2<Z<0.8)

= P(0.8)- P(-3.2)

= 0.7881−0.0007

=0.7874

2. What is the probability that a randomly selected bag weighs more than 4.8 kg?

= P(X>4.8)

$=P(\frac{X-\mu}{\sigma} > \frac{4.8-4.76}{0.05})$

= P(Z > 0.8)

= 1- P(Z ≤ 0.8)

= 1- 0.7881

= 0.2119

3. What is the probability that a randomly selected bag weighs less than 4.6 kg?

= P(X<4.6)

$=P(\frac{X-\mu}{\sigma} < \frac{4.6-4.76}{0.05})$

= P(Z < -3.2)

= 0.0007

2.6 Descriptive and inferential statistics:

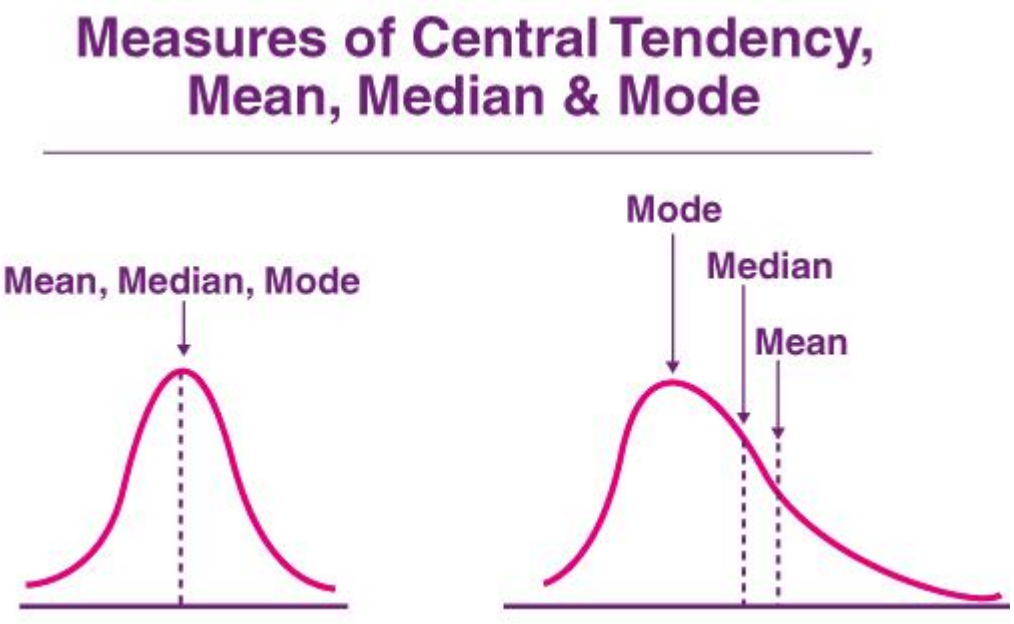
Descriptive statistics is a branch of statistics focused on summarizing and describing the key features of a dataset. It provides insights into the data's structure, spread, and central tendencies without drawing conclusions or making predictions.

Major component of Descriptive statistics:

1. Measures of Central Tendency(Mean, Median, Mode)
2. Measures of Variability(Range, Variance, Standard Deviation, Interquartile Range)
3. Shape of Distribution(Skewness, Kurtosis)
4. Frequency Distribution (Histograms, Boxplots, Piecharts, Barcharts)

Assignment: Difference between descriptive and inferential statistics.

Measure of Central Tendency: Measures of central tendency is a single value that attempts to describe a set of data by identifying the central position within the set of data. The mean, median and mode are 3 ways of finding the average.



1. Mean/Average: Mean is the most commonly used measure of central tendency. It actually represents the average of the given collection of data. It is applicable for both continuous and discrete data.

In an individual series, the mean is the arithmetic average of all data values.

$$\text{Mean}(\bar{X}) = \frac{\sum X}{N}$$

where, X represents each individual data value, and N is the total number of values in the dataset.

Example: If the data values are 5, 8, 10, 12, and 15:

$$\text{Mean} = \frac{5+8+10+12+15}{5} = \frac{50}{5} = 10$$

In a discrete series, each value X is associated with a frequency f (how many times that value occurs).

$$\text{Mean}(\bar{X}) = \frac{\sum(f \cdot X)}{\sum f}$$

where, f is the frequency of each value X and $\sum$f is the total sum of frequencies.

Example: Given the following data:

Value (X)	Frequency (f)
3	2
5	3
7	4

To find the mean:

$$\text{Mean} = \frac{(3 \times 2) + (5 \times 3) + (7 \times 4)}{2+3+4} = \frac{6+15+28}{9} = \frac{49}{9} \approx 5.44$$

In a continuous series, the data is grouped into class intervals, and each class interval has a frequency. The mean is calculated using the midpoints of these intervals.

$$\text{Mean}(\bar{X}) = \frac{\sum(f \cdot m)}{\sum f}$$

where, f is the frequency of each class, m is the midpoint of each class interval, and $\sum f$ is the total sum of frequencies.

Example: Given the following data:

Class Interval	Frequency (f)	Midpoint (m)	$f \cdot m$
10–20	3	15	45
20–30	5	25	125
30–40	2	35	70

To find the mean:

$$\text{Mean} = \frac{(3 \times 15) + (5 \times 25) + (2 \times 35)}{3 + 5 + 2} = \frac{45 + 125 + 70}{10} = \frac{240}{10} = 24$$

2. Median: The median is the middle value of a dataset that divides it into two equal halves.

In an individual series, we simply need to arrange data in ascending order. if the total number of values N is odd, the median is the middle value. If N is even, the median is the average of the two middle values.

$$\text{If } N \text{ is odd: Median} = X_{\left(\frac{N+1}{2}\right)}$$

$$\text{If } N \text{ is even: Median} = \frac{X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}}{2}$$

Odd: For data values 3, 8, 12, 15, and 20, the median is 12 (the middle value).

Even: For data values 6, 8, 10, and 14, the median is $\frac{8+10}{2} = 9$.

In a discrete series, each value has a frequency associated with it. To find the median, we calculate the cumulative frequencies to identify the middle position.

Steps:

* Arrange the values and frequencies in ascending order.
* Calculate the cumulative frequencies.
* Find the position of the median using $\frac{N}{2}$, where N is the total frequency.
* Identify the smallest value for which the cumulative frequency is greater than or equal to $\frac{N}{2}$. This value is the median.

Example Data:

Value (X)	Frequency (f)	Cumulative Frequency
10	3	3
20	5	8
30	2	10
40	1	11

Steps:

1. **Total Frequency** $N = 11$.
2. **Median Position** = $\frac{N}{2} = \frac{11}{2} = 5.5$.
3. Identify the value where the cumulative frequency is greater than or equal to 5.5. The cumulative frequency first exceeds 5.5 at 20 (where cumulative frequency = 8).

Median: The median is 20 for this discrete series.

For a continuous series, data is grouped into class intervals, so we use a formula involving the cumulative frequency of the classes.

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

Where:

- L = lower limit of the median class,
- N = total frequency,
- F = cumulative frequency of the class before the median class,
- f = frequency of the median class,
- h = width of the median class interval.

Example:

Class Interval	Frequency (f)	Cumulative Frequency
10–20	5	5
20–30	8	13
30–40	7	20

- Total frequency $N = 20$; $\frac{N}{2} = 10$.
- The median class is 20–30 because the cumulative frequency first exceeds 10 here.
- Applying the formula:
  - $L = 20, F = 5, f = 8, h = 10$.

$$\text{Median} = 20 + \left(\frac{10 - 5}{8} \right) \times 10 = 20 + \frac{5}{8} \times 10 = 20 + 6.25 = 26.25$$

3. Mode: Mode is the term appearing maximum time in data set i.e. term that has the highest frequency.

In an individual series, we simply need to count the frequency of each value in dataset. the one with maximum frequency is the Mode of the datset.
A dataset can be unimodal (one mode), bimodal (two modes), or multimodal (more than two modes).

Suppose we have the values 3, 7, 7, 10, and 15. Then the Mode = 7.
Suppose we have the values 4, 5, 8, 8, 10, 10, and 12. Modes = 8 and 10.
Suppose we have the values 2, 5, 9, and 14. There is no mode.

In a discrete series, each value X is associated with a frequency f, representing the count of occurrences.

Example:

Value (X)	Frequency (f)
5	3
10	5
15	5
20	2

Here, 10 and 15 both have the highest frequency (5), so this series is **bimodal with modes = 10 and 15.**

For a continuous series, data is grouped into class intervals. The mode is calculated by identifying the modal class — the class interval with the highest frequency. Then, we apply a formula to approximate the mode.

Formula:

$$\text{Mode} = L + \left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right) \times h$$

where:

- L = lower limit of the modal class,
- f_m = frequency of the modal class,
- f_1 = frequency of the class before the modal class,
- f_2 = frequency of the class after the modal class,
- h = class interval width.

Example:

Class Interval	Frequency (f)
10–20	6
20–30	15
30–40	10
40–50	4

- **Modal class** = 20–30 (highest frequency = 15)
- $L = 20,\ f_m = 15,\ f_1 = 6,\ f_2 = 10,\ h = 10$

Applying the formula:

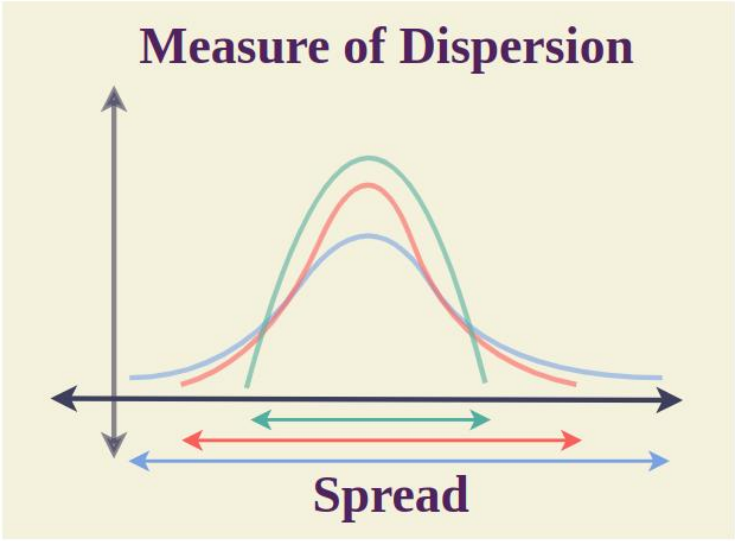
$$\text{Mode} = 20 + \left(\frac{15 - 6}{2 \times 15 - 6 - 10}\right) \times 10$$

$$= 20 + \left(\frac{9}{30 - 16}\right) \times 10$$

$$= 20 + \left(\frac{9}{14}\right) \times 10$$

$$= 20 + 6.43 = 26.43$$

Measures of Dispersion/Variability: Measures of Dispersion/Variability describe the spread, scatter, or variation of data points in a dataset. They indicate how much individual values differ from each other and from the central value (mean, median, or mode) of the data. The primary purpose of these measures is to quantify the degree of spread in the data, giving insight into data consistency, reliability, and comparison across datasets.



This is of two types:

1. Absolute measure of dispersion: The measures of dispersion which are expressed in terms of original units of a data are termed as Absolute measure of dispersion. It include Range, Quartile Deviation, Mean Deviation, Standard Deviation.

2. Relative measure of dispersion: These measures express dispersion as a ratio, percentage, or proportion, making it easier to compare variability across datasets that have different scales, units, or means. It includes Coefficient of Range, Coefficient of Quartile Deviation, Coefficient of Mean Deviation,Coefficient of Variation.

Range: The range is the simplest measure of dispersion. It tells us how spread out the values in a dataset are by calculating the difference between the highest and lowest values.

Formula:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Example:

- Data: 5, 8, 10, 12, 15
- Range = $15 - 5 = 10$

Interpretation: A larger range indicates more variability, while a smaller range indicates that data points are closer together.

For Descrete and Continuous Series:

$$\text{Range} = \text{Upper Limit of the Last Class Interval} - \text{Lower Limit of First Class Interval}$$

Quartile Deviation: The Quartile Deviation (QD), also known as the semi-interquartile range, measures the spread of the middle 50% of data. It is based on the interquartile range (IQR), which is the difference between the first and third quartiles (Q1 and Q3).

Interpretation:Quartile Deviation gives an indication of the data's spread around the median and is less affected by extreme values, making it useful for skewed distributions.

Mean Deviation (MD): The Mean Deviation (MD), also known as Mean Absolute Deviation, is the average of the absolute differences between each data point and the mean or median or mode of the dataset. MD helps in understanding the average spread

of the data.

Interpretation: MD gives an idea of the average distance of each value from the mean, providing insight into overall variability.

Standard Deviation(SD): Standard Deviation (SD) is a key measure of dispersion that shows how spread out data points are around the mean. It is calculated as the square root of the variance, providing an idea of the average distance of each data point from the mean.

Interpretation: A low SD means that the data points are close to the mean, indicating low variability. A high SD means that the data points are spread out over a larger range, indicating high variability.

Quartile: Self study(For Individual, Discrete, and Continuous Series)

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

$$Quartile\ Deviation = \frac{Q_3-Q_1}{2}$$

Where,

Q_3 = Upper Quartile (Size of $3[\frac{N+1}{4}]^{th}$ item)

Q_1 = Lower Quartile (Size of $[\frac{N+1}{4}]^{th}$ item)

$$Coefficient\ of\ Quartile\ Deviation = \frac{Q_3-Q_1}{Q_3+Q_1}$$

Where,

Q_3 = Upper Quartile (Size of $3[\frac{N+1}{4}]^{th}$ item)

Q_1 = Lower Quartile (Size of $[\frac{N+1}{4}]^{th}$ item)

More Examples on: https://easynotes4u.com/calculation-of-range-quartile-deviation-coefficient-measures-of-dispersion-absolute-relative-statistics/

Individual Series

$$Mean\ Deviation\ from\ Mean\ (MD_{\bar{X}}) = \frac{\sum|X-\bar{X}|}{N} = \frac{\sum|D|}{N}$$

$$Mean\ Deviation\ from\ Median\ (MD_{Me}) = \frac{\sum|X-M_e|}{N} = \frac{\sum|D|}{N}$$

Discrete Series

$$Mean\ Deviation\ from\ Mean\ (MD_{\bar{X}}) = \frac{\sum f|X-\bar{X}|}{N} = \frac{\sum f|D|}{N}$$

$$Mean\ Deviation\ from\ Median\ (MD_{Me}) = \frac{\sum f|X-M_e|}{N} = \frac{\sum f|D|}{N}$$

Continuous Series

$$Mean\ Deviation\ from\ Mean\ (MD_{\bar{X}}) = \frac{\sum f|m-\bar{X}|}{N} = \frac{\sum f|D|}{N}$$

$$Mean\ Deviation\ from\ Median\ (MD_{Me}) = \frac{\sum f|m-M_e|}{N} = \frac{\sum f|D|}{N}$$

Similarly for Mode.

For reference: https://easynotes4u.com/calculation-of-mean-deviation-in-individual-discrete-continuous-series-statistics/

$$\text{Coefficient of MD form mean} = \frac{\text{MD from mean}}{\text{Mean}}$$

$$\text{Coefficient of MD form median} = \frac{\text{MD from median}}{\text{Median}}$$

$$\text{Coefficient of MD form mode} = \frac{\text{MD from mode}}{\text{Mode}}$$

For Individual Series:

The standard deviation of the set of n observations is given by

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \quad \text{or} \quad \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

Where $\bar{x}$ is the arithmetic mean of the given observations.

For Discrete or Continuous Series:

For a discrete and continuous frequency distribution, the standard deviation is given by

$$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

Where $\bar{x}$ is the arithmetic mean and N is the total number of observations of the given data. In the case of a continuous distribution, x is taken as the mid-value of the corresponding class.

$$\text{Coefficient of S.D.} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

For individual data,

$$\text{Variance } (\sigma^2) = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

For discrete data,

$$\text{Variance } (\sigma^2) = \frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2$$

Coefficient of Variance (C.V.)

The Coefficient of Variance (C.V.) is a measure of dispersion equal to the standard deviation of a sample divided by the mean. It is dimensionless and not dependent on the units or scale in which the observations are made. It is often expressed as a percentage.

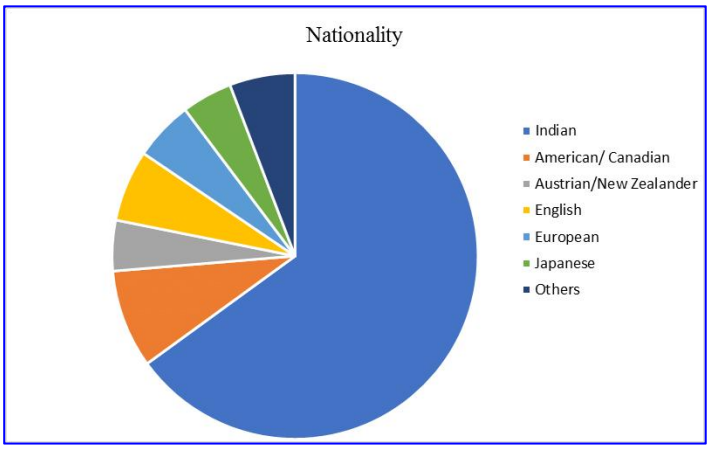
∴ The coefficient of variance (C.V.) is given by the formula,

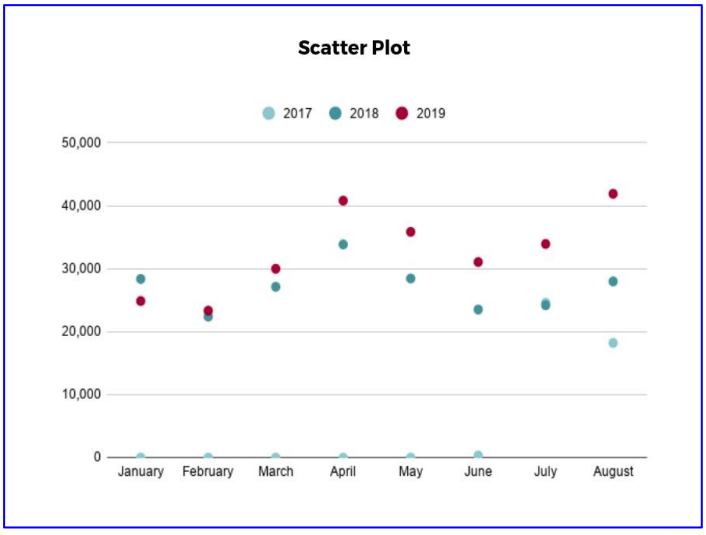
$$C.V. = \frac{\sigma}{\bar{x}} \times 100\%$$

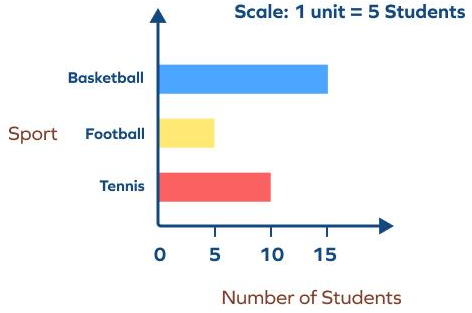
for examples: https://www.10mathproblems.com/2020/05/standard-deviation.html

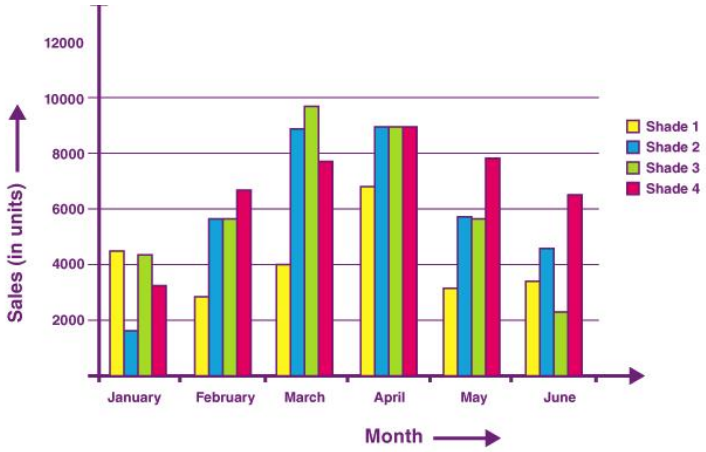
For more: https://easynotes4u.com/category/statistics/

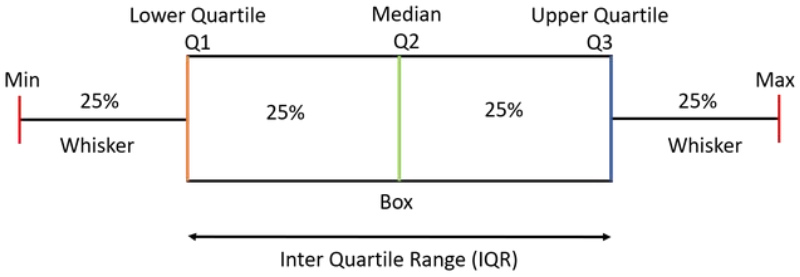
Data Describing with graphs:

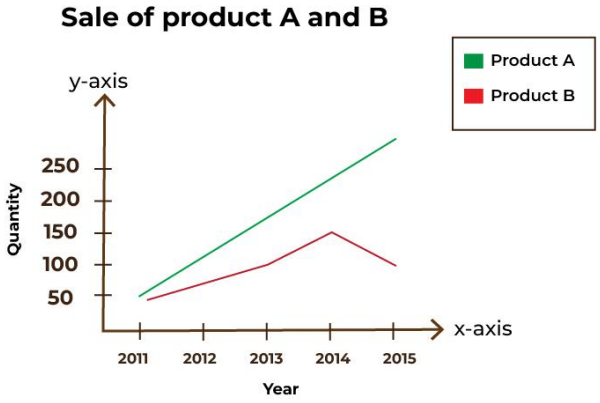












For more: https://www.10mathproblems.com/p/statistics.html

Inferential Statistics:

Inferential statistics is a branch of statistics that involves using sample data to make inferences or draw conclusions about a larger population. It is used to estimate population parameters, such as means, variances, or proportions, based on a subset of the data.

Inferential statistics comprises several techniques for drawing conclusions. Here are some common types:

1. Hypothesis Testing (z-test, t-test, f-test)
2. Confidence Intervals
3. Regression Analysis
4. Analysis of Variance (ANOVA)
5. Chi-Square Tests

Population: A large group of individuals and objects whose characteristics are being studied is called a population.

Sample: A small part of the population whcih is selested in random to evaluate the approximate characteristics of the population is called as sample..and the process is called sampling.

Parameters: The statistical measures computes form the popualtion is called as parameters. They describe certain properties fo polualtion. In other word, paramter is the function of population. i.e paramter = F(Population)

Statistics: The statistical measures computes form the samles is called as statistics. They describe certain properties fo Samples. In other word, paramter is the function of samples. i.e statistics= F(samples)

S.N.	Parameters	Statistics
1.	Population size = N	Sample size n
2.	Population mean = $\mu = \frac{\Sigma X}{N}$	Sample mean = $\bar{X} = \frac{\Sigma X}{n}$
3.	Population S.D. = $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	Sample S.D. = $S = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n-1}}$
4.	Population variance = σ^2	Sample variance = S^2
5.	Population proportion = $P = \frac{X}{N}$	Sample proportion = $\hat{p} = \frac{x}{n}$

2.7 Central limit theorem and sample distribution concepts

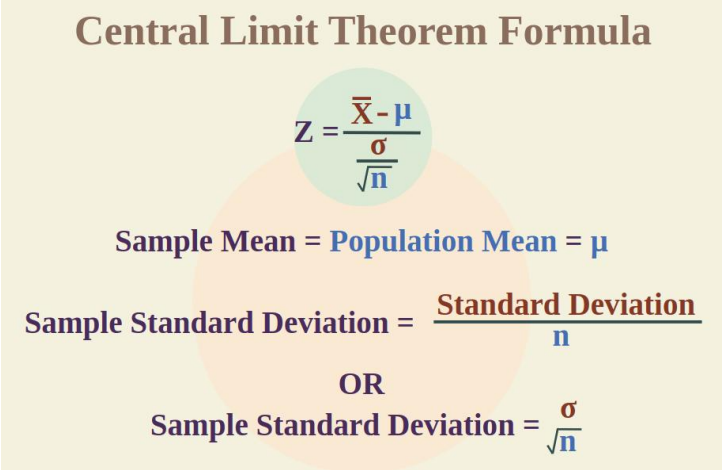
Central Limit Theorem:

It states that "For a sufficiently large sample size, the distribution of the sample mean will approximate a normal distribution, regardless of the shape of the population distribution, provided that the population has a finite mean and variance".

Key Points of the CLT:

1. The sample size must be large enough, typically $n \geqslant 30$.
2. The random variables in the samples are indepenedent and identiaclly distributed.
3. The sample is drawn from the population with a finite mean and variance.

If x1,x2,x3,⋯,xn is a random sample of size nn taken from any population with mean (μ) and variance ($\sigma 2$) , then the sampling distribution of the sample mean $(\bar{X})$ will approximate a normal distribution with mean(μ) and variance $\frac{\sigma 2}{n}$, provided that the sample size is sufficiently large.



We can use CLT to construct confidence intervals, perform hyothesis tests, and make prediction about population mean based on the sample data.

Sampling Distribution of the Sample Mean:

The sampling distribution of the sample mean is the distribution of the means of all possible samples of a given size from a population.

It has its own mean (equal to the population mean, μ) and a reduced standard deviation (called the standard error).

The Standard Error (SE) of the mean is the standard deviation of the sampling distribution of the sample mean, and it measures how much the sample mean is expected to vary from the actual population mean.

Formula:

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation, and n is the sample size.

A smaller SE indicates that the sample mean is closer to the population mean, improving the reliability of sample-based estimates. As the sample size increases, SE decreases, making estimates more precise.

1. Normal Approximation to the Sampling Distribution

When we draw random samples from any population, the distribution of sample means tends to approximate a **normal distribution** (according to the CLT) as the sample size n increases, even if the population itself is not normally distributed.

Key Assumptions for Normal Approximation:

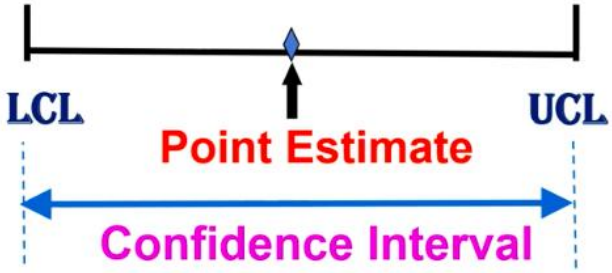
- The sample size n should be sufficiently large. Commonly, $n \geq 30$ is used as a rule of thumb, though the sample size might need to be larger for populations that are heavily skewed.
- The population from which the sample is drawn should have a finite mean (μ) and variance (σ^2).
- The **sampling distribution of the sample mean** will have:
  - A mean equal to the population mean: $\mu_{\bar{X}} = \mu$
  - A standard deviation (known as **standard error**) equal to $\frac{\sigma}{\sqrt{n}}$

As the sample size increases, the distribution of sample means becomes more tightly clustered around the population mean, and the shape approaches a normal distribution.

The process of estimating population parameters by using sample data is called estimation.

Point Estimate: If a single value calcualted from sample data is used to estimate population parameter then the process is called a point estimate.

Interval Estimate: If an interval or range of values calculated from sample data is used to estimate population parameter then the process is called a Interval estimate/Confidence Interval.



The formula for the confidence interval depends on the type of data and the sample size, specifically whether you are using the Z-distribution (for large sample sizes or known population standard deviation) or the t-distribution (for small sample sizes and unknown population standard deviation).

1. Confidence Interval for the Mean (Known Population Standard Deviation)

When the **population standard deviation** σ is known and the sample size n is large ($n \geq 30$), we use the **Z-distribution**.

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

Where:

- $\bar{X}$ = sample mean
- $Z_{\alpha/2}$ = Z-value corresponding to the desired confidence level (for example, for 95% confidence, $Z_{\alpha/2} \approx 1.96$)
- σ = population standard deviation
- n = sample size
- α = significance level (e.g., for a 95% confidence level, $\alpha = 0.05$)

2. Confidence Interval for the Mean (Unknown Population Standard Deviation)

When the **population standard deviation** σ is unknown and the sample size n is small ($n < 30$), we use the **t-distribution**.

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

Where:

- $\bar{X}$ = sample mean
- $t_{\alpha/2,\, df}$ = t-value corresponding to the desired confidence level and degrees of freedom $df = n - 1$
- s = sample standard deviation
- n = sample size
- α = significance level (e.g., for 95% confidence, $\alpha = 0.05$)
- df = degrees of freedom (for the t-distribution, $df = n - 1$)

a. Confidence Interval for the Mean (Known Standard Deviation)

Problem: A sample of 100 students has a mean score of 75 with a population standard deviation of 10. Calculate the 95% confidence interval for the mean score of all students.

- $\bar{X} = 75$
- $\sigma = 10$
- $n = 100$
- $Z_{\alpha/2} = 1.96$ (for 95% confidence)

$$CI = 75 \pm 1.96 \cdot \frac{10}{\sqrt{100}}$$

$$CI = 75 \pm 1.96 \cdot 1 = 75 \pm 1.96$$

$$CI = (73.04, 76.96)$$

So, the 95% confidence interval for the population mean score is **(73.04, 76.96)**.

b. Confidence Interval for the Mean (Unknown Standard Deviation)

Problem: A sample of 25 students has a mean score of 75 with a sample standard deviation of 10. Calculate the 95% confidence interval for the mean score of all students.

- $\bar{X} = 75$
- $s = 10$
- $n = 25$
- $\alpha = 0.05$ (for 95% confidence)
- Degrees of freedom $df = 25 - 1 = 24$
- The t-value for $\alpha/2 = 0.025$ and $df = 24$ is approximately **2.064** (from t-distribution tables).

$$CI = 75 \pm 2.064 \cdot \frac{10}{\sqrt{25}}$$

$$CI = 75 \pm 2.064 \cdot 2 = 75 \pm 4.128$$

$$CI = (70.872, 79.128)$$

So, the 95% confidence interval for the population mean score is **(70.872, 79.128)**.

Hypothesis testing procedures: Tests about the mean of a normal population:

When performing hypothesis testing for the mean of a normal population, the goal is to assess whether a sample mean provides sufficient evidence to infer something about the population mean. This is commonly done using either the Z-testor T-test, depending on whether the population standard deviation is known or unknown and the sample size.

Steps for Hypothesis Testing about Population Mean

1. State the Hypothesis:

a. Null Hypothesis (H_0): This represents the claim that the population mean is equal to a specific value. H_0: $\mu = \mu_0$

b. Alternative Hypothesis (H_1): This represents the claim that the population mean is different from the hypothesized value (two-tailed test), or greater than/less than the hypothesized value (one-tailed test).

Two-tailed test: H_1: $\mu \neq \mu_0$

One-tailed test:

Left-tailed: H1:$\mu<\mu_0$

Right-tailed: H1:$\mu>\mu_0$ Where μ_0 is the hypothesized population mean.

2. Select the Significance Level (α):

It is the max value of probability of rejecting null hypothesis when it is true. There can be two types of errors in testing of hypothesis.

Type I error: Rejecting null hypothesis when it is true. Represented by α.

Type II error: Accepting null hypothesis when it is false. Represented by β.

True State of H_0	Decision: Accept H_0 (Fail to Reject H_0)	Decision: Reject H_0
H_0 is True	**Correct Decision** (True Negative)	**Type I Error** (False Positive)
H_0 is False	**Type II Error** (False Negative)	**Correct Decision** (True Positive)

Explanation

- H_0 **is True and Accept** H_0: This is a **Correct Decision** (True Negative) — you correctly fail to reject H_0.
- H_0 **is True and Reject** H_0: This is a **Type I Error** (False Positive) — you incorrectly reject H_0 when it is actually true.
- H_0 **is False and Accept** H_0: This is a **Type II Error** (False Negative) — you fail to reject H_0 when it is actually false.
- H_0 **is False and Reject** H_0: This is a **Correct Decision** (True Positive) — you correctly reject H_0 when it is false.

3. Choose the Appropriate Test:

Z-test: If the population standard deviation (σ) is known and the sample size is large ($n \geqslant 30$).

T-test: If the population standard deviation (σ) is unknown and the sample size is small ($n<30n$).

- **Z-test** (when σ is known):

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- **T-test** (when σ is unknown):

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{X}$ = sample mean
- μ_0 = hypothesized population mean
- σ = population standard deviation (for Z-test)
- s = sample standard deviation (for T-test)
- n = sample size

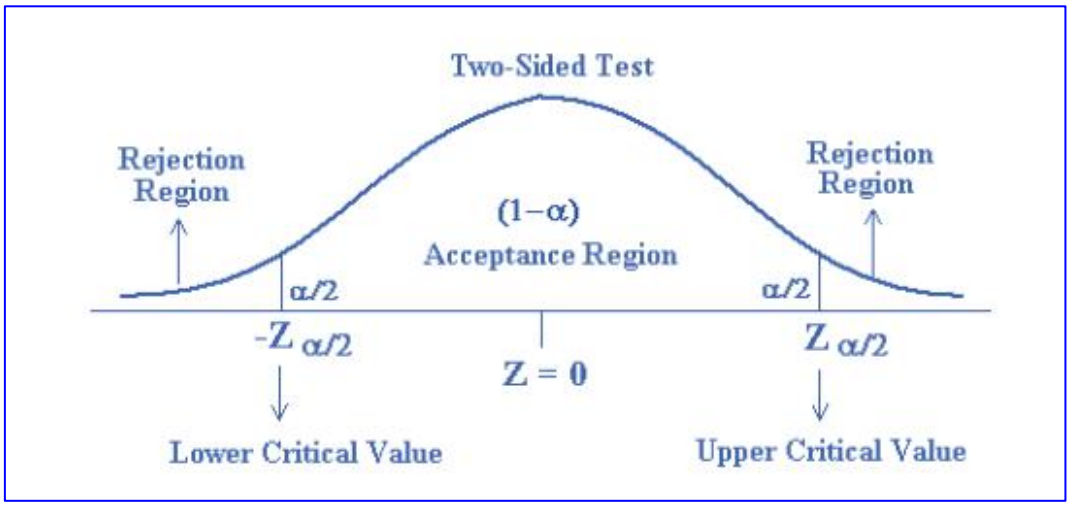
4. Determine Critical value:

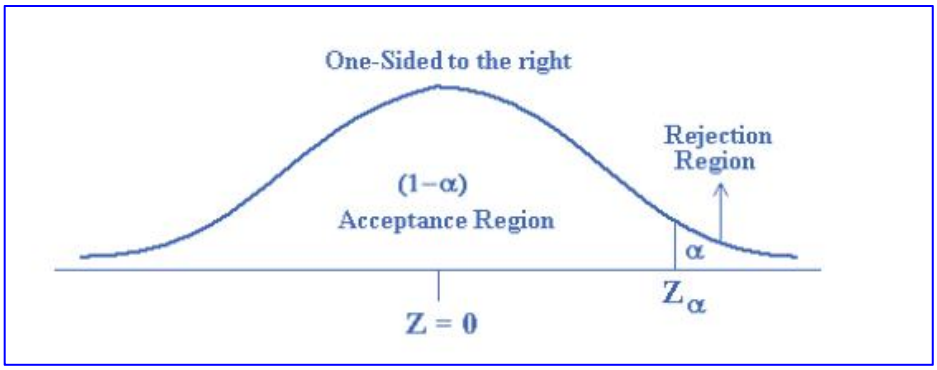
Using the formula on step 3 , we calculate value of test statistics =($Z_{calculated}$) or ($T_{calculated}$).

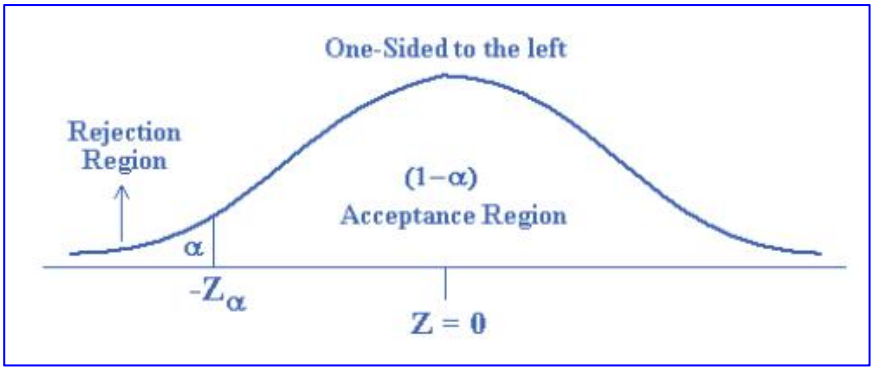
For Z tests, By using standard Z-tables based on the chosen α and the type of test (one-tailed or two-tailed) the Critical value of Z and named as Z_{table}.

For T tests, the critical t-value is found using t-distribution tables with the degrees of freedom df=n−1.

5. Make a Decision:







- In One tailed test(Right): The Null hypothesis (H_0) is rejected when $Z_{calculated} > Z_{\alpha}$. Else accepted.
- In One Tailed test (Left): The Null hypothesis (H_0) is rejected when $Z_{calculated} < -Z_{\alpha}$ Else accepted.
- In Two tailed test: The Null hypothesis (H_0) is rejected when $Z_{calculated} > Z_{\alpha/2}$ and $Z_{calculated} < -Z_{\alpha/2}$. Else accepted ($-Z_{\alpha/2} < Z < Z_{\alpha/2}$).

Example:

10. **A sample of 15 beams of a grand complex are given below in feet 3.**

12.8 9.8, 10.2, 10.0, 12.0, 10.5, 8.9, 12.2, 10.8, 9.0, 11.2, 12.1, 10.1, 8.8, 10.6

A civil engineer claimed that volume of beams are greater than 10 feet test whether his claim is right or wrong. [Use $\alpha = 1\%$]. [2075 Baishakh]

1. Null hypothesis:

   $H_0 : \mu = 10$ feet^3. The volume of beams are equal to 10 feet^3.

2. Alternate hypothesis:

   $H_1: \mu > 10$ feet^3. The volume of beams are greater than 10 feet^3.

3. Test statistic:

   $n = 15$

$$\bar{X} = \frac{12.8 + 9.8 + 10.2 + 10 + 12 + 10.5 + 8.9 + 12.2 + 10.8 + 9 + 11.2 + 12.1 + 10.1 + 8.8 + 10.6}{15}$$

$$= 10.6$$

$$S = \sqrt{\frac{(12.8-10.6)^2 + (9.8-10.6)^2 + 0.4^2 + 0.6^2 + 1.4^2 + 0.1^2 + 1.7^2 + 1.6^2 + 0.2^2 + 1.6^2 + 0.6^2 + 1.5^2 + 0.5^2 + 1.8^2 + 0^2}{15-1}}$$

$$= 1.257$$

As $n < 30$ and population standard deviation (σ) is not known. So, we use t-statistic.

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

$$= \frac{10.6 - 10}{\frac{1.257}{\sqrt{15}}} = 1.8487$$

4. For $\alpha = 0.01$ and $dt = n - 1 = 15 - 1 = 14$, the critical value is

   $t_{0.01,14} = 2.624$

5. **Decision:**

Since t_{cal} (1.8487) < t_{table} (2.624), We accept H_0. Thus the civil engineer's claim that the volume is greater than 10 feet^3 is wrong.

Ztest concerning Two Mean: (Hypothesis testing)

Let there be two sample size of n1 and n2 and (n1,n2>30) , sample SD of σ1 and σ2 respectively.

1. State Hypotheses:

Null Hypothesis (H0): μ1- μ2 = δ

Alternative Hypothesis (H1): μ1- μ2≠δ, μ1 - μ2>δ, or μ1- μ2<δ.

2. Choose Significance Level (α)

3. Compute Test Statistic: $Z_{calculated}$ using the formula

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups,
- δ is the hypothesized difference between the population means (often $\delta = 0$ when testing for equality),
- σ_1 and σ_2 are the known population standard deviations of the two groups,
- n_1 and n_2 are the sample sizes of the two groups.

4. Find critical value (Z_{table}) using Ztable. and Make decision based on Z_{table} and $Z_{calculated}$.

T test concerning Two Mean: (Hypothesis testing)

Let there be two sample size of n1 and n2 and (n1,n2<30) , Sample means of $\overline{x1}$, $\overline{x2}$ and sample SD of S1 and S2 respectively. δ is the difference of population means.

1. State Hypotheses:

Null Hypothesis (H0): μ1- μ2 = δ

Alternative Hypothesis (H1): μ1- μ2≠δ, μ1 - μ2>δ, or μ1- μ2<δ.

2. Choose Significance Level (α)

3. Compute Test Statistic: Tcalculated using the formula

Condition 1: if population variance is unknown and unequal.

If we assume that the variances of the two populations are **not equal**, we use Welch's t-test, which adjusts the formula and degrees of freedom:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The **degrees of freedom** are approximated as:

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

To calculate s_1 and s_2, we use the formula for the **sample standard deviation**:

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

Condition 2: if population variance is unknown and equal.

When the variances of the two samples are assumed to be equal, we use the **pooled standard deviation** s_p:

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

Then, the t-test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$df = n_1 + n_2 - 2$$

4. Find critical value (Ttable) using Ttable and degree of freedom(df). and Make decision based on Ttable and Tcalculated.

3. The mean consumption of food grains among 400 sampled middle class consumers is 380 grams per day per person with standard deviation of 120 gram. A similar sample survey of 600 working class consumers gave a mean of 410 grams with standard deviation of 80 grams can you justify that the classes consumes the same quantity of food grain, using 5% level of significance. [2076 Bhadra]

Solution:

Given,

$n_1 = 400$

$\bar{X}_1 = 380$ grams/day/person

$\sigma_1 = 120$ grams

$n_2 = 600$

$\bar{X}_2 = 410$ grams/day/person

$\sigma_2 = 80$ grams

1. Null hypothesis: $H_0: \mu_1 = \mu_2$

   The classes consume the same quantity of food grams.

2. Alternative hypothesis: $H_1: \mu_1 \neq \mu_2$

   The classes do not consume the same quantity of food grains.

3. Test statistic:

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

$$Z = \frac{380 - 410}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{380 - 410}{\sqrt{\frac{120^2}{400} + \frac{80^2}{600}}}$$

$$= -4.39155$$

4. Critical value:

   As it is two tailed test So,

$$Z_{\frac{0.05}{2}} = Z_{0.025} = \pm 1.96$$

5. Decision:

   Since $|Z_{cal}| > |Z_{table}|$, we reject null hypothesis and accept alternative hypothesis i.e. the consumes different quantity.

37. **10 samples of standard cement had shown an average weight percent calcium with mean 90 and variance 25.15 samples of the lead dropped cement has an average weight percent of 87 with variance 16. If the weight percent in population is normally distributed with same variance, test the hypothesis for equality of mean at $\alpha = 0.01$ level of significance.** [2070 Magh]

Solution:

$n_1 = 10, \bar{X}_1 = 90, S_1^2 = 25$

$n_2 = 15, \bar{X}_2 = 87, S_2^2 = 16$

1. Null hypothesis: $H_0: \mu_1 = \mu_2$

   The mean of the two cement types are same.

2. Alternate hypothesis: $H_1: \mu_1 \neq \mu_2$

   Since, $n < 30$ and population mean (σ) is not given we use t-score

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

$$= \frac{9 \times 25 + 14 \times 16}{10 + 15 - 2} = 19.52$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{sp^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(90 - 87)}{\sqrt{19.52 \times \left(\frac{1}{10} + \frac{1}{15}\right)}} = 1.6632$$

4. Critical value:

   $df = n_1 + n_2 - 2$

   $= 10 + 15 - 2$

   $= 23$ two-tailed test

   $t = 2.807$

5. Decision:

   Since, $t_{cal} < t_{table}$, we accept H_0 and the mean are same.

2.10 ANOVA

Analysis of Variance (ANOVA) is a statistical method used to compare the means of three or more independent groups to determine if there are any statistically significant differences among them. It helps in understanding whether the variation in a dataset is due to the differences between group means or if it is due to random variation within the groups.

Types of ANOVA:

One way ANOVA and Two Way ANOVA

1. One-Way ANOVA: (Comparing Means Across Multiple Groups)
One-way ANOVA (Analysis of Variance) is a statistical test used to determine whether there are statistically significant differences in the means of three or more independent groups. It helps to test the null hypothesis that all group means are equal.

Assumptions of One-Way ANOVA:

* Independence of Observations: The samples in each group must be independent of each other.

* Normality: The data within each group should follow a normal distribution.

* Homogeneity of Variances: The variances of the groups should be approximately equal.

2. Two-Way ANOVA:
Two-way ANOVA is an extension of one-way ANOVA used to analyze the effect of two independent categorical variables (factors) on a dependent variable. It also evaluates whether there is an interaction effect between the two factors.

Example: Used in experiments where two factors are studied simultaneously, such as
You want to investigate how teaching methods (Factor A) and study duration (Factor B) affect students' test scores(dependent variable).

One Way Anova: (Fisher's test)

Step 1: Set up hypothesis.

Null Hypothesis: $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$ (H_0)

Alternative Hypothesis: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_n$ (H_1)

We reject H_0 if $F_{ratio} > F_{\alpha}(v_1, v_2)$ where,

$v_1 = t - 1$, and

$v_2 = n - t$

Step 2: Calculation:

Treatment (t)	parameter (r)				
T_1	y_{11},	y_{12},	y_{13},	y_{1j}, ...	y_{1r}
T_2	y_{21},	y_{22},	y_{23},	y_{2j}, ...	y_{2r}
T_3	⋮	⋮	⋮	⋮	⋮
⋮					
T_i	y_{i1},	y_{i2},	y_{i3} ...	y_{ij} ...	y_{ir}
T_t	y_{t1},	y_{t2},	y_{t3} ...	y_{tj} ...	y_{tr}

Sum	mean
$y_{1\cdot}$	$\bar{y}_1$
$y_{2\cdot}$	$\bar{y}_2$
⋮	⋮
$y_{t\cdot}$	$\bar{y}_t$

Total (grand total) $= y_{\cdot\cdot}$ grand mean (GM) $(\bar{\bar{y}})$

(a) Correction factor = $C = \frac{y_{..}^2}{n}$

(b) Sum of square of total (SST) = $\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - C$

(c) Sum of square of treatment (SST_r) = $\frac{1}{r}\sum_{i=1}^{t} y_{i.}^2 - C$

(d) Sum of square of error (SSE) = $SST - SST_r$

(e) Treatment mean square $(MST)_r = \frac{SST_r}{t-1}$

(f) Error mean square (MSE) = $\frac{SSE}{n-t}$

(g) F ratio = $\frac{MST_r}{MSE}$

In tabular form,

Source	degree of freedom	sum of square	mean square	F ratio
Treatment	$t-1$	SST_r	MST_r	$F_{ratio} = \frac{MST_r}{MSE}$
error	$n-t$	SSE	MSE	
Total	$n-1$	SST		

3. Level of significance (α)

4. Critical value
Obtain $F_{critical}$ / F_{table} from the F table based on α

5. Decision:
If $F_{ratio} \leq F_{table}$, H_0 is accepted.
If $F_{ratio} > F_{table}$, H_0 is rejected.
ie. H_1 is accepted.

Example:

product (t)	sales (r)
A	14, 13, 9, 15, 11, 13, 14, 11
B	10, 12, 9, 7, 11, 8, 12, 9, 10, 13, 9, 10
C	11, 5, 9, 10, 6, 8, 8, 7

Construct Anova table and test the hypothesis.

Solution:

step 1: Set up hypothesis:

Null hypothesis: $\mu_1 = \mu_2 = \mu_3$ (H_0)

Alternative hypothesis: $\mu_1 \neq \mu_2 \neq \mu_3$ (H_1)

step 2: Calculating:

product (t)	sales (r)	Sum	mean
A	14, 13, 9, 15, 11, 13, 14, 11	100 ($y_{1.}$)	100/8 = 12.5
B	10, 12, 9, 7, 11, 8, 12, 9, 10, 13, 9, 10	120 ($y_{2.}$)	120/12 = 10
C	11, 5, 9, 10, 6, 8, 8, 7	64 ($y_{3.}$)	64/8 = 8
Total:		$y_{..}$ = 284	$\bar{y}$ = 10.16667

Correction factor (C) = $\frac{y_{..}^2}{n} = \frac{284^2}{28} = 2880.571429$

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{r} (y_{ij})^2 - C = 14^2 + 13^2 + \ldots + 8^2 + 7^2 - 2880.57$$

$$= 3052 - 2880.57$$

$$= 171.428571$$

$$SST_r = \frac{1}{r}\sum_{i=1}^{t} y_{i.}^2 - C$$

$$= \frac{1}{r}\sum_{i=1}^{3} y_{i.}^2 - C$$

$$= \left(\frac{100^2}{8} + \frac{120^2}{12} + \frac{64^2}{8}\right) - 2280.57$$

$$= 81.428571$$

$$SSE = SST - SST_r$$

$$= 171.428571 - 81.428571$$

$$= 90$$

$$MST_r = \frac{SST_r}{t-1} = \frac{81.428571}{3-1} = 40.714285$$

$$MSE = \frac{SSE}{n-t} = \frac{90}{28-3} = 3.6$$

$$F_{ratio} = \frac{MST_r}{MSE} = \frac{40.714285}{3.6} = 11.30952375$$

3. level of significance (α)

$\alpha = 0.01$ i.e. (99% confidence)

$F_\alpha(\theta_1, \theta_2)$ where, $\theta_1 = t-1 = 3-1 = 2$

$= F_{0.01}(2, 25)$ $\theta_2 = n-t = 28-3 = 25$

From table,

we get, $F_{table} = 5.57$

4. Decision;

Here, $F_{ratio} > F_{table}$

$11.30 > 5.57$

So, we reject H_0 and accept H_1. //