DMC6101 MATRICES, PROBABILITY AND STATISTICS

OBJECTIVES:

- To provide methods for understanding the consistency and solving the equation as well as for finding the Eigenvalues and Eigenvectors of square matrix.
- To provide foundation on Applied Probability
- To introduce the concepts of correlation and regression of random variables
- To use various statistical techniques in Application problems
- To introduce the concept of Design of Experiments for data analysis

UNIT - I MATRICES AND EIGENVALUE PROBLEMS

Matrices - Rank of a Matrix - Consistency of a system of linear equations - Solution of the matrix equation $\Delta x=b$ - Row - reduced Echelon Form - Eigenvalues and Eigenvectors - Properties - Cayley - Hamilton Theorem - Inverse of a matrix.

UNIT - II PROBABILITY AND RANDOM VARIABLES

Probability - Axioms of Probability - Conditional Probability - Addition and multiplication laws of Probability - Baye's theorem - Random Variables - Discrete and continuous random variables - Probability mass function and Probability density functions - Cumulative distribution function - Moments and variance of random variables - Properties - Binomial, Poisson, Geometric, Uniform, Exponential, Normal distributions and their properties.

UNIT - III TWO-DIMENTIONAL RANDOM VARIABLES

15

Joint probability distributions - Marginal and conditional probability distributions - Covariance - Correlation - Linear regression lines - Regression curves - Transform of random variables - Central limit theorem (for independent identically random variables).

UNIT - IV TESTING OF HYPOTHESIS

Sampling distributions - Tests based on small and large samples - Normal, Student's t, Chi-square and F distributions for testing of mean, variance and proportion and testing of difference of means variances and proportions - Tests for independence of attributes and goodness of fit.

UNIT - V DESIGN OF EXPERIMENTS

Analysis of variance - Completely randomized design - Random block design (One-way and Two-way classifications) - Latin square design - 2^2 Factorial design.

OUTCOMES:

After the completion of the course the student will be able to

- Test the consistency and solve system of linear equations as well as find the Eigenvalues and Eigenvector.
- Apply the Probability axioms as well as rules and the distribution of discrete and continuous ideas in solving real world problems.
- Apply the concepts of correlation and regression of random variables in solving application problems.
- Use statistical techniques in testing hypothesis on data analysis.
- Use the appropriate statistical technique of design of experiments in data analysis.

REFERENCE BOOKS:

1. B.S. Grewal, Higher Engineering Mathematics, Khanna Publishers, 43rd Edition, New Delhi, 2015.
2. R.K. Jain and S.R.K Iyenger, Advanced Engineering Mathematics, Narosa Publishing House, New Delhi, 2002.
3. Devore, J.L, Probability and Statistics for Engineering and Sciences, Cengage Learning, 8th Edition, New Delhi, 2014.
4. Miller and M. Miller, Mathematical Statistics, Pearson Education Inc., Asia 7th Edition, New Delhi, 2011.
5. Richard Johnson, Miller and Freund's Probability and Statistics for Engineer, Prentice Hall of India Private Ltd., 8th Edition, New Delhi, 2011.

MATRICES, PROBABILITY AND STATISTICS
SCHEME OF LESSONS

UNIT -I
MATRICES AND EIGENVALUE PROBLEMS

CONTENTS

UNIT -I
MATRICES AND EIGENVALUE PROBLEMS

Learning Objectives

In this chapter a student has to learn the

- Rank of a matrix and methods finding these.
- Consistency of System of equations and Solving these.
- Finding Eigen values and Eigen vectors of square matrix
- Inverse of a matrix using Cayley-Hamilton Theorem.

Learning Outcomes

Upon completion of the lesson, students are able to demonstrate a good understanding of:

- How to find rank of a matrix
- How to identify consistency of system of equations
- Solving system of equations using row-echelon form.
- To finding eigen values and eigen vectors of a matrix.
- Finding inverse and higher positive power of a matrix.

OVERVIEW

In this lesson, you are going to study about the Matrices and learn how to find rank of matrices. These basic concepts will help you to understand the concept of consistency of a system of equations. Learn how to find the eigen values and eigen vector of a matrix. Application of Cayley-Hamilton theorem.

1.1 Introduction

Definition:

A rectangular array of numbers is called a matrix. We shall mostly be concerned with matrices having real numbers as entries. The horizontal arrays of a matrix are called its ROWS and the vertical arrays are called its COLUMNS. A matrix having m rows and n columns is said to have the order m X n.

A matrix A of ORDER m X n can be represented in the following form:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{1j} & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{2j} & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & a_{mj} & a_{mn} \end{bmatrix}_{n\,x\,m}$$

where a_{ij} is the entry at the intersection of the i^{th} row and jth column.

In a more concise manner, we also denote the matrix A by $[a_{ij}]$ by suppressing its order.

Special Matrices

Definition:

1. A matrix in which each entry is zero is called a zero-matrix, denoted by $\mathbf{0}$. For example,

$$\mathbf{0}_{2\times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \mathbf{0}_{2\times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

2. A matrix for which the number of rows equals the number of columns, is called a square matrix. So, if A is an n X n matrix then A is said to have order n .
3. In a square matrix, $A = [a_{ij}]$ of order n , the entries are called $a_{11}, a_{22}, a_{33}, \dots, a_{nn}$ the diagonal entries and form the principal diagonal of A.
4. A square matrix $A = [a_{ij}]$ is said to be a diagonal matrix if $a_{ij} = 0$ for $i \neq j$ In other words, the non-zero entries appear only on the principal diagonal. For example, the zero matrix 0_n and $\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$ are a few diagonal matrices.

   A diagonal matrix D of order n with the diagonal entries $d_1, d_2, \dots, d_n$ is denoted by $D = \text{diag}(d_1, d_2, \dots, d_n)$ If $d_i = d$ for all i =1,2,3,..,n then the diagonal matrix D is called a **scalar matrix**.
5. A diagonal matrix A of order n is called an IDENTITY MATRIX if $d_i = 1$ for all i =1,2,3,..,n . This matrix is denoted by I_n .
   For example,

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

   The subscript n is suppressed in case the order is clear from the context or if no confusion arises.
6. A square matrix $A = [a_{ij}]$ is said to be an upper triangular matrix if $a_{ij} = 0$ for $i > j$
   A square matrix $A = [a_{ij}]$ is said to be a lower triangular matrix if $a_{ij} = 0$ for $i < j$
   A square matrix A is said to be triangular if it is an upper or a lower triangular matrix.
   For example
   $\begin{bmatrix} 2 & 1 & 4 \\ 0 & 3 & -1 \\ 0 & 0 & -2 \end{bmatrix}$ is an upper triangular matrix. An upper triangular matrix will be represented by

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

Operations on Matrices

Definition: (Transpose of a Matrix) The transpose of an m x n matrix $A = [a_{ij}]$ is defined as the n x m matrix $B = [b_{ij}]$ with $a_{ij} = b_{ij}$ for $1 \leq\leq i \leq m$ and $1 \leq\leq i \leq n$ The transpose of A is denote*d by* A^T.

That is, by the transpose of an m x n matrix A we mean a matrix of order n x m having the rows of A as its columns and the columns of A as its rows.

For example, if

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

Then $A^T = \begin{bmatrix} 1 & 0 \\ 4 & 1 \\ 5 & 2 \end{bmatrix}$

Thus, the transpose of a row vector is a column vector and vice-versa.

Definition (Addition of Matrices) let $A = [a_{ij}]$ and $B = [b_{ij}]$ be are two m x n matrices. Then the sum A +B is defined to be the matrix $C = [c_{ij}]$ with$c_{ij} = a_{ij} + b_{ij}$

Note that, we define the sum of two matrices only when the order of the two matrices are same.

Definition (Multiplying a Scalar to a Matrix) Let $A = [a_{ij}]$ be an m x n matrix. Then for any element kϵR we define $kA = [ka_{ij}]$

For example, if

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

and k =5 then

$$5A = \begin{bmatrix} 5 & 20 & 25 \\ 0 & 5 & 10 \end{bmatrix}.$$

Theorem: Let A, B and C be matrices of order m x n and let k, lϵR Then

1. A + B = B +A (Commutativity).
2. (A+B)+C = A + (B+C) (associativity)
3. K(lA)=(kl)A
4. (k+l)A = kA + lA

Definition (Additive Inverse) Let A *be an m x n matrix.*

1. Then there exists a matrix B with A+B = 0 The matrix B is called the additive inverse of A and is denoted by -A =(-1)A
2. Also, for the matrix 0_{mxn}, A + 0 = 0 +A =A Hence, the matrix 0_{mxn}, is called the additive identity.

1.2 RANK OF A MATRIX`

Minor of a matrix:

Let A be any given matrix of order m x n. The determinant of any submatrix of a square order is called minor of the matrix A.

We observe that, if 'r' denotes the order of a minor of a matrix of order m x n then $1 \leq r \leq m$ if $m<n$ and $1 \leq r \leq n$ if $n<m$

Definition: Rank of a matrix:

A number 'r' is called rank of a matrix of order mxn if there is atleast one minor of the matrix which is of order r whose value is non-zero and all the minors of order greater than, 'r' will be zero. The rank of a matrix is denoted by 'ρ'

e.g.(i) Let

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 4 & 1 \\ 3 & 5 & 7 \end{bmatrix}$$

Consider e.g. Let

$$A_1 = \begin{vmatrix} 1 & 0 \\ 2 & 4 \end{vmatrix} = 4,\ A_2 = \begin{vmatrix} 0 & 2 \\ 4 & 1 \end{vmatrix} = -8 \text{ etc.}$$

$$A_3 = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 4 & 1 \\ 3 & 5 & 7 \end{bmatrix} = 1(23)+2(-2) = 19 \neq 0$$

$\therefore$ Rank of A= 3

(ii) $A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 0 & -1 & -1 \end{bmatrix}$

Here,

$$A_1 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 0 & -1 & -1 \end{bmatrix} = 1(1)-1(-1)+2(-1) = 0$$

$$A_2 = \begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix} = 1 \neq 0$$

Thus minor of order 3 is zero and atleast one minor of order 2 is non-zero

$\therefore$ Rank of A = 2.

Some results:

(i) Rank of null matrix is always zero
(ii) Rank of any non-zero matrix is always greater than or equal to 1.
(iii) Rank of transpose of matrix A is always equal to rank of A.
(iv) Rank of product of two matrices cannot exceed the rank of both of the matrices
(v) Rank of a matrix remains unleased by elementary transformations.

Elementary Transformations:

Following changes made in the elements of any matrix are called elementary transactions.

(i) Interchanging any two rows (or columns) .
(ii) Multiplying all the elements of any row (or column) by a non-zero real number.
(iii) Adding non-zero scalar multitudes of all the elements of any row (or columns) into the corresponding elements of any another row (or column).

Definition:

A matrix is in row-echelon form if

i. Any row consisting of all zeros is at the bottom of the matrix.
ii. For all non-zero rows the leading entry must be a one. This is called the pivot.
iii. In consecutive rows the pivot in the lower row appears to the right of the pivot in the higher row.

1.3 Determining Consistency

Rouch´e-Capelli Theorem States: The system Ax = B admits solutions (it is consistent) if and only if rank(A) = rank(A|B).

Moreover, if the system is consistent, the number of degrees of freedom is equal to n − rank(A), where n is the number of unknowns of the system.

Given the linear system AX = B and the augmented matrix (A|B).

(i) if rank(A) = rank(A|B) = the number of rows in x, then the system has a unique solution.
(ii) If rank(A) = rank(A|B) < the number of rows in x, then the system has ∞-many solutions
(iii) If rank(A) < rank(A|B), then the system is inconsistent.

Example:1 Discuss the consistency of the system of equations

$$2x+3y-4z=-2$$
$$x-y+3z=4$$
$$3x+2y-z=-5$$

Solution: In the matrix form

$$\begin{bmatrix} 2 & 3 & -4 \\ 1 & -1 & 3 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \\ -5 \end{bmatrix}$$

Consider an Agumental matrix

$$[A:D]=\begin{bmatrix} 2 & 3 & -4 & : & -2 \\ 1 & -1 & 3 & : & 4 \\ 3 & 2 & -1 & : & -5 \end{bmatrix} \quad \begin{matrix} R_2 \to R_2 - \frac{1}{2} R_1 \\ R_3 \to R_3 - \frac{3}{2} R_1 \end{matrix}$$

$$[A:D]=\begin{bmatrix}2 & 3 & -4 & : & -2\\ 0 & -\frac{5}{2} & 5 & : & 5\\ 0 & -\frac{5}{2} & 5 & : & -2\end{bmatrix} \qquad R_2 \to R_3 - R_2$$

$$[A:D]=\begin{bmatrix}2 & 3 & -4 & : & -2\\ 0 & -\frac{5}{2} & 5 & : & 5\\ 0 & 0 & 5 & : & -7\end{bmatrix}$$

$$\therefore\ \rho(AD)=3$$

$$\rho(A)=2$$

$$\therefore\ \rho(AD)\neq\rho(A)$$

$\therefore$ The system is inconsistent and it has no solution.

Example:2

Discuss the consistency of the system of equations

$3x+y+2z=3$

$2x-3y-z=-3$

$x+2y+z=4$

Solution: In the matrix form,

$$\begin{bmatrix}3 & 1 & 2\\ 2 & -3 & -1\\ 1 & 2 & 1\end{bmatrix}\begin{bmatrix}x\\ y\\ z\end{bmatrix}=\begin{bmatrix}3\\ -3\\ 4\end{bmatrix}$$

$A \qquad X = D$

Now we join matrices A and D

Consider

$$[A:D]=\begin{bmatrix}3 & 1 & 2 & : & 3\\ 2 & -3 & -1 & : & -3\\ 1 & 2 & 1 & : & 4\end{bmatrix} \qquad \begin{matrix}R_2 \to R_2 - 2R_1\\ R_3 \to R_3 - 3R_1\end{matrix}$$

$$[A:D]=\begin{bmatrix}1 & 2 & 1 & : & 4\\ 2 & -7 & -3 & : & -11\\ 0 & -5 & -1 & : & -9\end{bmatrix} \qquad R_3 \to R_3 - \frac{5}{7}R_2$$

$$[A:D]=\begin{bmatrix}1 & 2 & 1 & : & 4\\ 0 & -7 & -3 & : & -11\\ 0 & 0 & \frac{8}{7} & : & -\frac{8}{7}\end{bmatrix} \quad \text{.....(1)}$$

This is in Echelon form

$$\therefore \rho(AD) = 3$$

$$\rho(A) = 3$$

$\therefore \rho\ (\mathrm{AD}) = \rho\ (\mathrm{A}) =$ Number of unknowns

$\therefore$ system is consist and has unique solution.

Step (2) : To find the solution we proceed as follows. At the end of the row transformation the value of z is calculated then values of y and the value of x in the last.

The matrix in example 1 in Echelon form can be written as,

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & -7 & -3 \\ 0 & 0 & 8/7 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ -11 \\ -8/7 \end{bmatrix}$$

$\therefore$ Expanding by R_3

$$\frac{8}{7}Z = -\frac{8}{7}$$

$\therefore$ z = -1

$\therefore$ expanding by R_2

$$-7y - 3z = -11$$

$$-7y - 3(-1) = -11$$

$$-7y + 3 = -11$$

$$+7y = +14$$

$$y = 2$$

exp*anding* by R_1

$$x + 2y + z = 4$$

$$x + 4 - 1 = 4$$

$\therefore$ x = 1

$\therefore$ x = 1, y = 2, z = -1

Example:3

Examine for consistency 5x+3y+7z =4; 3x+26y+2z =9; 7x+2y+10z =5

Solution:

Step (1) : In the matrix form

$$\begin{bmatrix} 5 & 3 & 7 \\ 3 & 26 & 2 \\ 7 & 2 & 10 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

$$A \qquad X = D$$

Consider

$$[A:D] = \begin{bmatrix} 5 & 3 & 7 & : & 4 \\ 3 & 26 & 2 & : & 9 \\ 7 & 2 & 10 & : & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

$$[A:D]=\begin{bmatrix}1 & 3/5 & 7/5 & : & 4/5\\ 3 & 26 & 2 & : & 9\\ 7 & 2 & 10 & : & 5\end{bmatrix} \qquad R_1 \to \frac{1}{5} R_1$$

$$[A:D]=\begin{bmatrix}1 & 3/5 & 7/5 & : & 4/5\\ 0 & 121/5 & -11/5 & : & 33/5\\ 0 & -11/5 & 1/5 & : & -3/5\end{bmatrix} \qquad \begin{matrix}R_2 \to R_2 - 3R_1\\ R_3 \to R_3 - 7R_1\end{matrix}$$

$$[A:D]=\begin{bmatrix}1 & 3/5 & 7/5 & : & 4/5\\ 0 & 121/5 & -11/5 & : & 33/5\\ 0 & 0 & 0 & : & 0\end{bmatrix} \qquad R_3 \to R_3 + \frac{1}{11} R_2$$

$\therefore \quad \rho(AD) = 2$

$\rho(A) = 2$

$\therefore \quad \rho(AD) = \rho(A) = 2 < 3 = Number$ of unknowns

The system is consistent and has infinitely many solutions.

1.4 Eigen Values and Eigen Vectors

Definition:

Let A = { a_{ij}} be a square matrix of order n, If there exists a non-zero column vector X and a scalar λ such that AX = λ X then the λ is called an eigen value of the matrix A and X is called the eigen vector cooresponding to the eigen value of λ.

Note:

The Equation | A - λ I| = o is called the characteristic equation of A. the solutions of characteristic equation is called eigen values(say $\lambda_1,\lambda_2,\lambda_3,\ldots.$) of A.

Corresponding to each value of λ, the above equation posses a non-zero solution X. X is called the eigen vector of A corresponding to the eigen values of A.

Corresponding to each value of λ, the above equation posses a non-zero solution X. X is called the eigen vector of A corresponding to the eigen values of A.

Formula to find the characteristic equation of 2 X 2 matrix:

λ^2 – (Sum of the main diagonal elements) λ + | A |= 0.

Formula to find the characteristic equation of 3X3 matrix:

λ^3 – (Sum of the main diagonal elements of A (or) Trace of A) λ^2 + (Sum of the minors of the main diagonal elements of A) λ - | A | = 0.

Properties of Eigen values:

1) A square matrix A and its transpose A^T have the same eigen values.

Let $A = (a_{ij})$, $i,j = 1,2,3$

The characteristic polynomial of A

is $|A - \lambda I| = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}$ …(1)

Then the characteristic polynomial of A^T is

$|A^T - \lambda I| = \begin{vmatrix} a_{11}-\lambda & a_{21} & a_{31} \\ a_{12} & a_{22}-\lambda & a_{32} \\ a_{13} & a_{23} & a_{33}-\lambda \end{vmatrix}$ …(2)

Determinant (2) can be obtained by changing rows into columns of determinant (1).

Therefore $|A - \lambda I| = |A^T - \lambda I|$

Therefore the characteristic equation of A and A^T is identical.

Therefore the eigen values of A and A^T are the same.

2) a) The sum of the eigen values of a matrix A is equal to the sum of the principal diagonal elements of A. (The sum of the principal diagonal elements is called the trace of the matrix)

b) The product of the eigen values of A is equal to $|A|$.

Proof of 2 .a)

The characteristic equation of an 3rd order matrix A may be written as

λ^3 – (Sum of the main diagonal elements of A (or) Trace of A) λ^2

+ (Sum of the minors of the main diagonal elements of A)λ - $|A| = 0$.

Let $\lambda_1,\lambda_2,\lambda_3$ be the eigen values of a matrix A. …(1)

Then (1) can be re written as $(\lambda-\lambda_1)(\lambda-\lambda_2)(\lambda-\lambda_3)=0$

That is $\lambda^3 - (\lambda_1+\lambda_2+\lambda_3)\lambda^2 + (\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3)\lambda - \lambda_1\lambda_2\lambda_3 = 0$ …(2)

Equating the coefficient of λ^2 in (1) and (2) , Then

Sum of the main diagonal elements of $A = \lambda_1+\lambda_2+\lambda_3$ = Trace of A.

Therefore The sum of the eigen values of a matrix A is equal to the sum of the principal diagonal elements of A.

Proof of 2. b)

Equating the constant coefficient in (1) and (2)

Then $|A| = \lambda_1\lambda_2\lambda_3$.

Therefore the product of the eigen values of A is equal to $|A|$

Note: If $|A| = 0$, i.e., A is a singular matrix, at least one of the eigen values of A is 0 and conversely.

3) If $\lambda_1,\lambda_2,\lambda_3,\ldots \lambda_n$ are the eigen values of a matrix A, then

i) $k\lambda_1,k\lambda_2,k\lambda_3,\ldots k\lambda_n$ are the eigen values of the matrix kA, where k is a non-zero scalar.

ii) $\lambda_1^P,\lambda_2^P,\lambda_3^P,\ldots \lambda_n^P$ are the eigen values of the matrix A^P, where P is a positive integer.

iii) $1/\lambda_1, 1/\lambda_2, 1/\lambda_3,\ldots 1/\lambda_n$ are the eigen values of the inverse matrix A^{-1}, provided λ_r not equal to 0 i.e., A is non-singular.

Proof of I) Let λ_r be an eigen value of A and X_r the corresponding eigen vector. Then , by definition $A X_r = \lambda_r X_r$ ....(1)

Multiplying both sides of (1) by k

$$kA X_r = k\lambda_r X_r \quad \ldots(2)$$

From (2) $k\lambda_r$ are the eigen values of the matrix kA and the corresponding eigen vector is the same as that of λ_r Namely (X_r).

Proof of ii)

Multiplying both sides of (1) by A

$$A^2 X_r = A (A X_r)= A(\lambda_r X_r) = \lambda_r^2 X_r \ .$$

Similarly $A^3 X_r =\lambda_r^3 X_r$ .and so on.

In general, $A^P X_r =\lambda_r^P X_r$. (3)

From (3) λ_r^P is the eigen values of the matrix A^P, with the corresponding eigen vector is the same as that of λ_r.Namely (X_r).

Proof of iii)

Multiplying both sides of (1) by A^{-1}.

$$A^{-1}(AX_r) = A^{-1}(\lambda_r X_r)$$

$$X_r = \lambda_r (A^{-1} X_r)$$

Therefore $A^{-1} X_r = \frac{1}{\lambda r} X_r \quad \ldots(4)$

From (4) $\frac{1}{\lambda r}$ is an eigen value of A^{-1} with the corresponding eigen vector is the same as that of λ_r .Namely (X_r).

Example:4

For the given matrix of order $3, |A| = 32$ and two of its Eigen values are 8 & 2. Find the sum of the Eigen values.

Solution:

Let the eigen values be $\lambda_1, \lambda_2, \lambda_3$.

Given $|A| = 32, \lambda_1 = 8 \ \& \ \lambda_2 = 2$.

Since $|A| = 32$

$\lambda_1\lambda_2\lambda_3$=32

i.e., $(8)(2)\lambda_3 = 32$

$\Longrightarrow \lambda_3 = 2$.

Sum of the eigen values $=\lambda_1+\lambda_2 + \lambda_3 = 8 + 2 + 2 = 12$.

Example:5

If the sum of the two eigen values and trace of a 3×3 matrix A are equal, find the value of $|A|$.

Solution: Let the eigen values be $\lambda_1, \lambda_2, \lambda_3$.

Given sum of the two eigen values = trace of A.

i.e., $\lambda_1+\lambda_2 = \lambda_1 + \lambda_2 + \lambda_3$

$\Longrightarrow \lambda_3 = 0$

$\therefore |A| = \lambda_1\lambda_2\lambda_3 = 0.$

Example:6

If 1 & 2 are the eigen values of a 2×2matrix A, then find the eigen values of A^2 and $A^{-1.}$

Solution:

Let $\lambda_1 = 1, \lambda_2$ =2 be the eigen values of A.

Then eigen values of A^2 are $\lambda_1{}^2$ & , $\lambda_2{}^2$ i.e, eigen values of A^2 are 1& 4.

Similarly the eigen values of A^{-1} are $\frac{1}{\lambda_1}$ & $\frac{1}{\lambda_2}$ i.e., eigen values of A^{-1} are1& $\frac{1}{4}$.

Example: 7

Given $A = \begin{pmatrix} -1 & 0 & 0 \\ 2 & -3 & 0 \\ 1 & 4 & 2 \end{pmatrix}$. Find the eigen values of A^T.

Solution:

Since A is a triangular matrix the eigen values of A are $-1, -3$ & 2.

By the property, eigen values of A & A^T are equal.

$\therefore$ the eigen values of A^T are $-1, -3$ & 2.

Example: 8

If the eigen values of A are 1,2 , 3 then what are the eigen values of Adj A.

Solution:

:

The eigen value are $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3.$

The eigen values of adjA are $\frac{|A|}{\lambda_1}, \frac{|A|}{\lambda_2}, \frac{|A|}{\lambda_3}$

$|A| = 1 \times 2 \times 3 = 6.$

$\therefore$ the eigen values of adjA are $\frac{6}{1}, \frac{6}{2}, \frac{6}{3} = 6,3,2.$

Example: 9

Check whether the matrix Bis orthogonal? JustifyB $= \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Solution:

Bis an orthogonal matrix if $BB^T = I.$

Consider $BB^T = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$

$\therefore$ Bis an orthogonal matrix.

Example: 10

If A & B are non-singular matrices then prove that AB & BA will have the same eigen values.

Solution:

Given A & B are non-singular matrices

Consider $AB = IAB = (B^{-1}B)(AB) = B^{-1}(BA)B$.

Hence AB & BA are similar matrices.

Since eigen values of similar matrices are same , AB & BA will have the same eigen values.

Example: 11

Find the constants a & b such that the matrix $\begin{pmatrix} a & 4 \\ 1 & b \end{pmatrix}$ has 3 & -2 as its eigen values.

Solution:

The eigen value are $\lambda_1 = 3 \& \lambda_2 = -2$

$\lambda_1 + \lambda_2 = a + b \text{ i.e.}, 3 - 2 = a + b$

$a + b = 1 \quad ----(1)$

$\lambda_1\lambda_2 = ab - 4 \text{ i.e.}, ab - 4 = -6$

$ab = -2 \quad ---(2)$

From (2) $b = \frac{-2}{a}$

Substituting $b = \frac{-2}{a}$ in (1)

$a - \frac{2}{a} = 1 \text{ i.e.}, a^2 - 2 = a$

$\Longrightarrow a^2 - a - 2 = 0 \text{ i.e.}, (a-2)(a+1) = 0$

$a = 2, -1$

If $a = 2 \implies b = -1$

If $a = -1 \implies b = 2.$

Example: 12

Find the eigen values and eigen vectors of $\begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix}$.

Solution:

Eigen Values:

The characteristic equation is given by $\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$

where S_1 = sum of the main diagonal elementt of the matrix $= 1 + 2 - 1 = 2$

S_2 = sum th minors of the main diagonal elements of the matrix

$$= \begin{vmatrix} 2 & -1 \\ 1 & -1 \end{vmatrix} + \begin{vmatrix} 1 & 4 \\ 2 & -1 \end{vmatrix} + \begin{vmatrix} 1 & -1 \\ 3 & 2 \end{vmatrix} = (-2+1) + (-1-8) + (2+3) = -5$$

S_3 = Derterimants of the matrix

$$= \begin{vmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{vmatrix} = 1(-2+1) + 1(-3+2) + 4(3-4) = -6$$

Thus the characteristic equation is $\lambda^3 - S_1\lambda^2 + S_2\lambda + S_3 = 0$

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

1	1	-2	-5	6
	0	1	-1	-6
	1	-1	-6	0

$\lambda = 1, \lambda^2 - \lambda - 6 = 0 \implies (\lambda - 3)(\lambda + 2) = 0 \text{ i.e.}, \lambda = 3, \lambda = -2.$

$\therefore \lambda = -2, 1, 3.$

Eigen Vectors :

The eigen vectors are given by $(A-\lambda I)X=0$.

i.e., $\begin{pmatrix}1-\lambda & -1 & 4\\ 3 & 2-\lambda & -1\\ 2 & 1 & -1-\lambda\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=0$ ---------- (1)

Case(i) $\lambda=-2$

Put $\lambda=-2$ in $equation\ (1)$

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

$\frac{x_1}{\begin{vmatrix}-1 & 4\\ 4 & -1\end{vmatrix}}=\frac{x_2}{-\begin{vmatrix}3 & 4\\ 3 & -1\end{vmatrix}}=\frac{x_3}{\begin{vmatrix}3 & -1\\ 3 & 4\end{vmatrix}}$ i.e., $\frac{x_1}{(1-16)}=\frac{x_2}{-(-3-12)}=\frac{x_3}{(12+3)}$

$\frac{x_1}{-15}=\frac{x_2}{15}=\frac{x_3}{15}$ i.e., $\frac{x_1}{-1}=\frac{x_2}{1}=\frac{x_3}{1}$

The eigen vector $X_1=\begin{pmatrix}-1\\ 1\\ 1\end{pmatrix}$

Case(ii) $\lambda=1$

Put $\lambda=1$ in (1)

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

$\frac{x_1}{\begin{vmatrix}-1 & 4\\ 1 & -1\end{vmatrix}}=\frac{x_2}{-\begin{vmatrix}0 & 4\\ 3 & -1\end{vmatrix}}=\frac{x_3}{\begin{vmatrix}0 & -1\\ 3 & 1\end{vmatrix}}$ i.e., $\frac{x_1}{(1-4)}=\frac{x_2}{-(0-12)}=\frac{x_3}{(0+3)}$

$\frac{x_1}{-3}=\frac{x_2}{12}=\frac{x_3}{3}$ i.e., $\frac{x_1}{-1}=\frac{x_2}{4}=\frac{x_3}{1}$

The eigen vector $X_2=\begin{pmatrix}-1\\ 4\\ 1\end{pmatrix}$.

Case(iii) $\lambda=3$

Put $\lambda=3$ in (1)

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

$\frac{x_1}{\begin{vmatrix}-1 & 4\\ -1 & -1\end{vmatrix}}=\frac{x_2}{-\begin{vmatrix}-2 & 4\\ 3 & -1\end{vmatrix}}=\frac{x_3}{\begin{vmatrix}-2 & -1\\ 3 & -1\end{vmatrix}}$ i.e., $\frac{x_1}{(1+4)}=\frac{x_2}{-(2-12)}=\frac{x_3}{(2+3)}$

$\frac{x_1}{5}=\frac{x_2}{10}=\frac{x_3}{5}$ i.e., $\frac{x_1}{1}=\frac{x_2}{2}=\frac{x_3}{1}$

The eigen vector $X_3=\begin{pmatrix}1\\ 2\\ 1\end{pmatrix}$.

Example: 13

Find the eigen values and eigen vectors of $\begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}$.

Solution:

Eigen Values :

The characteristic equation is given by $\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$

where $S_1 = 2 + 3 + 2 = 7$

$$S_2 = \begin{vmatrix} 3 & 1 \\ 2 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 2 \\ 1 & 3 \end{vmatrix} = (6-2) + (4-1) + (6-2) = 11$$

$$S_3 = \begin{vmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{vmatrix} = 2(6-2) - 2(2-1) + 1(2-3) = 5$$

Thus the characteristic equation is $\lambda^3 - 7\lambda^2 + 11\lambda - 5 = 0$

1		1	-7	11	-5
		0	1	-6	5
		1	-6	5	0

$\lambda = 1, \lambda^2 - 6\lambda + 5 = 0 \implies (\lambda - 1)(\lambda - 5) = 0$ i.e., $\lambda = 1, \lambda = 5$.

$\therefore \lambda = 1,1,5.$

Eigen Vectors :

The eigen vectors are given by $(A - \lambda I)X = 0$.

i.e., $\begin{pmatrix} 2-\lambda & 2 & 1 \\ 1 & 3-\lambda & 1 \\ 1 & 2 & 2-\lambda \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$ ---------- (1)

Case(i) $\lambda = 5$

Put $\lambda = 5$ in (1)

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

$$\frac{x_1}{\begin{vmatrix} 2 & 1 \\ -2 & 1 \end{vmatrix}} = \frac{x_2}{-\begin{vmatrix} -3 & 1 \\ 1 & 1 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} -3 & 2 \\ 1 & -2 \end{vmatrix}} \text{ i.e., } \frac{x_1}{(2+2)} = \frac{x_2}{-(-3-1)} = \frac{x_3}{(6-2)}$$

$\frac{x_1}{4} = \frac{x_2}{4} = \frac{x_3}{4}$ i.e., $\frac{x_1}{1} = \frac{x_2}{1} = \frac{x_3}{1}$

The eigen vector $X_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

Case(ii) $\lambda = 1$

Put $\lambda = 1$in (1)

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

$x_1 + 2x_2 + x_3 = 0$

Let $x_1 = 0$ we have $2x_2 + x_3 = 0$ i.e., $2x_2 = -x_3$

$\frac{x_2}{-1} = \frac{x_3}{2}$

The eigen vector $X_2 = \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}$.

Let $x_2 = 0$ we have $x_1 + x_3 = 0$ i.e., $x_1 = -x_3$

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

The eigen vector $X_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$.

1.5 Cayley- Hamilton Theorem

Statement: Every square matrix satisfies its own characteristic equation.

Note: When λ is replaced by A in the characteristic equation, the constant term 'c' should be replaced by 'cI' to get the result of Cayley-Hamilton theorem, Where I is the unit matrix of order n, also 0 in the right side of the above equation is a null matrix of order n

1.5.1 Application of Cayley-Hamilton theorem

(i) If A is non-singular, we can get A^{-1}

(ii) Higher positive integral powers of A can be computed, if we know powers of A of lower degree.

Example: 14

If 2,3 are the eigen values of a square matrix A of order 2, express A^2 in terms of A & I.

Solution:

Characteristic equation of A is $\lambda^2 - (\text{sum of the eigen values})\lambda + (\text{product of the eigen values}) = 0$

i.e., $\lambda^2 - (2+3)\lambda + (2 \times 3) = 0$

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

By Cayley Hamilton theorem, $A^2 - 5A + 6I = 0$

$$A^2 = 5A - 6I$$

Example: 15

Verify that the matrix $A = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}$ satisfies its characteristic equation and hence find A^4.

Solution:

Eigen Values:

The characteristic equation is given by $\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$

where $S_1 = 2 + 2 + 2 = 6$

$$S_2 = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 2 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = (4-1) + (4-2) + (4-1) = 8$$

$$S_3 = \begin{vmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{vmatrix} = 2(4-1) + 1(-2+1) + 2(1-2) = 3$$

Thus the characteristic equation is $\lambda^3 - 6\lambda^2 + 8\lambda - 3 = 0$.

Verification:

To verify Cayley – Hamilton theorem we should show that

$A^3 - 6A^2 + 8A - 3I = 0$ ----------(1)

To find A^2 & A^3 :

$$A^2 = A.A = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}\begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 7 & -6 & 9 \\ -5 & 6 & -6 \\ 5 & -5 & 7 \end{pmatrix}$$

$$A^3 = A^2.A = \begin{pmatrix} 7 & -6 & 9 \\ -5 & 6 & -6 \\ 5 & -5 & 7 \end{pmatrix}\begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 29 & -28 & 38 \\ -22 & 23 & -28 \\ 22 & -22 & 29 \end{pmatrix}$$

Consider,

$A^3 - 6A^2 + 8A - 3I =$

$$= \begin{pmatrix} 29 & -28 & 38 \\ -22 & 23 & -28 \\ 22 & -22 & 29 \end{pmatrix} - 6\begin{pmatrix} 7 & -6 & 9 \\ -5 & 6 & -6 \\ 5 & -5 & 7 \end{pmatrix} + 8\begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} - 3\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

To find A^4:

From (1)

$A^3 = 6A^2 - 8A + 3I$ ---------- (2)

Pre – Multiplying (2) by A

$A^4 = 6A^3 - 8A^2 + 3A$

$$= 6\begin{pmatrix} 29 & -28 & 38 \\ -22 & 23 & -28 \\ 22 & -22 & 29 \end{pmatrix} - 8\begin{pmatrix} 7 & -6 & 9 \\ -5 & 6 & -6 \\ 5 & -5 & 7 \end{pmatrix} + 3\begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}$$

$$A^4 = \begin{pmatrix} 124 & -123 & 162 \\ -95 & 96 & -123 \\ 95 & -95 & 124 \end{pmatrix}.$$

Example: 16

Using Cayley – Hamilton theorem find A^{-1} & A^4 for the matrix

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

Solution:

Eigen Values :

The characteristic equation is given by $\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$

where $S_1 = 2 + 2 + 2 = 6$

$$S_2 = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = (4-1) + (4-1) + (4-1) = 9$$

$$S_3 = \begin{vmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{vmatrix} = 2(4-1) + 1(-2+1) + 1(1-2) = 4$$

Thus the characteristic equation is$\lambda^3 - 6\lambda^2 + 9\lambda - 4 = 0$.

By Cayley – Hamilton theorem replacing λby A

$A^3 - 6A^2 + 9A - 4I = 0$ ----------(1)

To find A^2 & A^3 :

$$A^2 = A.A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & -5 & 5 \\ -5 & 6 & -5 \\ 5 & -5 & 6 \end{pmatrix}$$

$$A^3 = A^2.A = \begin{pmatrix} 6 & -5 & 5 \\ -5 & 6 & -5 \\ 5 & -5 & 6 \end{pmatrix}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 22 & -21 & 21 \\ -21 & 22 & -21 \\ 21 & -21 & 22 \end{pmatrix}$$

To find A^4:

From (1)

$A^3 = 6A^2 - 9A + 4I$ ----------(2)

Pre – Multiplying (2) by A

$$A^4 = 6A^3 - 9A^2 + 4A$$

$$= 6\begin{pmatrix} 22 & -21 & 21 \\ -21 & 22 & -21 \\ 21 & -21 & 22 \end{pmatrix} - 9\begin{pmatrix} 6 & -5 & 5 \\ -5 & 6 & -5 \\ 5 & -5 & 6 \end{pmatrix} + 4\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 86 & -85 & 85 \\ -85 & 86 & -85 \\ 85 & -85 & 86 \end{pmatrix}.$$

To find A^{-1}:

From (1)

$4I = A^3 - 6A^2 + 9A$ ----------(3)

Post-Multiplying (3) by A^{-1}

$$4A^{-1} = A^2 - 6A + 9I$$

$$A^{-1} = \frac{1}{4}(A^2 - 6A + 9I)$$

$$= \frac{1}{4}\left[\begin{pmatrix} 6 & -5 & 5 \\ -5 & 6 & -5 \\ 5 & -5 & 6 \end{pmatrix} - 6\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix} + 9\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]$$

$$A^{-1} = \frac{1}{4}\begin{pmatrix} 3 & 1 & -1 \\ 1 & 3 & 1 \\ -1 & 1 & 3 \end{pmatrix}.$$

Example: 17

Use Cayley – Hamilton theorem to find the value of the matrix given by

$(A^8 - 5A^7 + 7A^6 - 3A^5 + A^4 - 5A^3 + 8A^2 - 2A + I)$ if the matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}.$$

Solution:

Eigen Values:

The characteristic equation is given by $\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$

where $S_1 = 2 + 1 + 2 = 5$

$$S_2 = \begin{vmatrix} 1 & 0 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = (2 - 0) + (4 - 1) + (2 - 0) = 7$$

$$S_3 = \begin{vmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{vmatrix} = 2(2 - 0) - 1(0 - 0) + 1(0 - 1) = 3$$

Thus the characteristic equation is$\lambda^3 - 5\lambda^2 + 7\lambda - 3 = 0$.

By Cayley – Hamilton theorem replacing λby A

$A^3 - 5A^2 + 7A - 3I = 0$ ------- (1)

Let the polynomial $\lambda^8 - 5\lambda^7 + 7\lambda^6 - 3\lambda^5 + \lambda^4 - 5\lambda^3 + 8\lambda^2 - 2\lambda + 1$ be divided by $\lambda^3 - 5\lambda^2 + 7\lambda - 3$.

$$
\begin{array}{l|l}
 & \lambda^5 + \lambda \\
\hline
\lambda^3 - 5\lambda^2 + 7\lambda - 3 & \lambda^8 - 5\lambda^7 + 7\lambda^6 - 3\lambda^5 + \lambda^4 - 5\lambda^3 + 8\lambda^2 - 2\lambda + 1 \\
 & (-)\quad \lambda^8 - 5\lambda^7 + 7\lambda^6 - 3\lambda^5 \\
\hline
 & \lambda^4 - 5\lambda^3 + 8\lambda^2 - 2\lambda \\
 & (-)\quad \lambda^4 - 5\lambda^3 + 7\lambda^2 - 3\lambda \\
\hline
 & \lambda^2 + \lambda + 1
\end{array}
$$

Thus

$\lambda^8 - 5\lambda^7 + 7\lambda^6 - 3\lambda^5 + \lambda^4 - 5\lambda^3 + 8\lambda^2 - 2\lambda + 1$

$= (\lambda^3 - 5\lambda^2 + 7\lambda - 3)(\lambda^5 + \lambda) + (\lambda^2 + \lambda + 1)$

Replacing λ by A

$A^8 - 5A^7 + 7A^6 - 3A^5 + A^4 - 5A^3 + 8A^2 - 2A + I$

$= (A^3 - 5A^2 + 7A - 3I)(A^5 + A) + (A^2 + A + I)$

$= A^2 + A + I$ (from (1)

To find A^2:

$$A^2 = A.A = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 4 & 4 \\ 0 & 1 & 0 \\ 4 & 4 & 5 \end{pmatrix}$$

$$A^8 - 5A^7 + 7A^6 - 3A^5 + A^4 - 5A^3 + 8A^2 - 2A + I = \begin{pmatrix} 5 & 4 & 4 \\ 0 & 1 & 0 \\ 4 & 4 & 5 \end{pmatrix} + \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 8 & 5 & 5 \\ 0 & 3 & 0 \\ 5 & 5 & 8 \end{pmatrix}.$$

$$\therefore A^8 - 5A^7 + 7A^6 - 3A^5 + A^4 - 5A^3 + 8A^2 - 2A + I = \begin{pmatrix} 8 & 5 & 5 \\ 0 & 3 & 0 \\ 5 & 5 & 8 \end{pmatrix}.$$

Example: 18

If the matrix $A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2+i & -1 & 0 & 0 \\ -3 & 2i & i & 0 \\ 4 & -i & 1 & -i \end{pmatrix}$ where$i = \sqrt{-1}$, then using Cayley- Hamilton theorem prove that$A^4 = I$.

Solution:

Since A is a triangular matrix the eigen values of A are $1, -1$, i & $-$ i.
The characteristic equation is given by $(\lambda - 1)(\lambda + 1)(\lambda - i)(\lambda + i) = 0$

$$(\lambda^2 - 1)(\lambda^2 + 1) = 0$$

$$\lambda^4 - 1 = 0$$

By Cayley - Hamilton theorem, $A^4 - I = 0$
i.e., $A^4 = I$.

Example: 19

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

Solution:

The characteristic equation of A is given by $\lambda^2 - (1 + 3)\lambda + (3 - 8) = 0$
i.e., $\lambda^2 - 4\lambda - 5 = 0$.
By Cayley Hamilton theorem $A^2 - 4A - 5I = 0$
Pre multiplying by A^{-1}we get $A - 4I - 5A^{-1} = 0$
i.e., $5A^{-1} = A - 4I$

$$A^{-1} = \frac{1}{5}(A - 4I) = \frac{1}{5}\left(\begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix} - \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right) = \frac{1}{5}\begin{pmatrix} -3 & 4 \\ 2 & -1 \end{pmatrix}$$

$$\therefore A^{-1} = \frac{1}{5}\begin{pmatrix} -3 & 4 \\ 2 & -1 \end{pmatrix}.$$

SUMMARY

In this chapter we have learn

- Types of matrices and operations of matrices
- Rank of a matrix
- Testing consistency of a system of equation
- Characteristics equation & its root
- Eigen vector which is corresponding to Eigen value
- Cayley Hamilton theorem & its application like Higher power of matrix & Inverse of matrix

Keywords

- **Rank of a matrix**: A number 'r' is called rank of a matrix of order mxn if there is atleast one minor of the matrix which is of order r whose value is non-zero and all the minors of order greater than, 'r' will be zero. The rank of a matrix is denoted by 'ρ'
- **Consistency of a system of equations**: The system Ax = B admits solutions (it is consistent) if and only if rank(A) = rank(A|B). Moreover if the system is consistent, the number of degrees of freedom is equal to n − rank(A), where n is the number of unknowns of the system
- **Eigen values and Eigen vectors:**

  Let A = { a_{ij}} be a square matrix of order n, If there exists a non-zero column vector X and a scalar λ such that AX = λ X then the λ is called an eigen value of the matrix A and X is called the eigen vector cooresponding to the eigen value of λ.

- **characteristic Equation: $|A - \lambda I| = 0$**

Formula to find the characteristic equation of 3X3 matrix:

λ^3 – (Sum of the main diagonal elements of A (or) Trace of A) λ^2 + (Sum of the minors of the main diagonal elements of A) λ - | A | = 0.

- **Cayley-Hamilton Theorem**: Every square matrix satisfies its own characteristic equation.

SELF-ASSESSMENT QUESTIONS

Short Answer Questions

1. Find the sum and product of the eigen values of $A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 0 & 3 \\ -2 & -1 & -3 \end{pmatrix}$.
2. Find the eigen values of A^2 & A^{-1} of $A = \begin{pmatrix} 2 & 5 & -1 \\ 0 & 3 & 2 \\ 0 & 0 & 4 \end{pmatrix}$.
3. Show that the matrix $A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ is orthogonal.
4. Prove that eigen values of $-3A^{-1}$ are the same as those of $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$.
5. One of the eigen values of $\begin{pmatrix} 7 & 4 & -4 \\ 4 & -8 & -1 \\ 4 & -1 & -8 \end{pmatrix}$ is -9. Find the other two eigen values.
6. Using Cayley – Hamilton theorem find the inverse of $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.
7. Write down the applications of Cayley - Hamilton theorem.

Long Answer Questions

8. Solve the system of equations:
   $x_1 + x_2 + x_3 = 3,\ x + 2x_2 + 3x_3 = 4,\ x_1 + 4x_2 + 9x_3 = 6$
9. Solve the system of equations:
   $x_1 - 4x_2 - x_3 = 3$
   $3x_1 + x_2 - 2x_3 = 7$
   $2x_1 - 3x_2 + x_3 = 10.$
10. Use Cayley - Hamilton theorem to find $A^4 - 4A^3 - 5A^2 + A + 2I$ when $A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}$.
11. If $\boldsymbol{\lambda}$ is an eigen value of a matrix A then show that $\lambda^n$is an eigen value of A^n.
12. Find the eigen values and eigen vectors of $\begin{pmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{pmatrix}$.
13. Find the eigen values and eigen vectors of $\begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}$
14. Verify Cayley – Hamilton theorem and hence find A^{-1} if
    $A = \begin{pmatrix} 1 & 2 & -2 \\ -1 & 3 & 0 \\ 0 & -2 & 1 \end{pmatrix}$
15. Use Cayley – Hamilton theorem to find the value of the matrix given by $(A^8 - 5A^7 + 7A^6 - 3A^5 + 8A^4 - 5A^3 + 8A^2 - 2A + I)$ if the matrix

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}.$$

16. Using Cayley – Hamilton theorem find A^{-1} & A^4 for the matrix
$$A = \begin{pmatrix} 1 & 3 & 7 \\ 4 & 2 & 3 \\ 1 & 2 & 1 \end{pmatrix}.$$
17. Find A^nusing Cayley – Hamilton theorem, taking$A = \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}$ and also find A^3.

FURTHER READINGS

1. B.S. Grewal, Higher Engineering Mathematics, Khanna Publishers, 43rd Edition, New Delhi, 2015.
2. R.K. Jain and S.R.K Iyenger, Advanced Engineering Mathematics, Narosa Publishing House, New Delhi, 2002.

UNIT II
PROBABILITY AND RANDOM VARIABLES

CONTENTS

Learning Objectives

In this chapter a student has to learn the

- Discrete and continuous random variable
- Expectations
- Moment Generating Function
- Solving problems on Binomial, Poisson Distribution, Geometric Distribution
- Solving problems on Uniform, Exponential Distribution and Normal Distribution

Learning Outcomes

Upon completion of this Unit, students are able to demonstrate a good understanding of:

- Difference between discrete and continuous random variables
- Finding moments about origin and arbitrary constants
- Moments generating function for Discrete and continuous distributions
- Solving real life problems using distributions

Overview

In this Unit, you are going to study about the random variables and learn how to find mean and variance. These basic concepts will help you to understand the concept of distribution function. Application of Distribution functions both in Discrete and Continuous cases.

2.1 Introduction

Rolling an ordinary six-sided die is a familiar example of a random experiment, an action for which all possible outcomes can be listed, but for which the actual outcome on any given trial of the experiment cannot be predicted with certainty. In such a situation we wish to assign to each outcome, such as rolling a two, a number, called the probability of the outcome, that indicates how likely it is that the outcome will occur. Similarly, we would like to assign a probability to any event, or collection of outcomes, such as rolling an even number, which indicates how likely it is that the event will occur if the experiment is performed. This section provides a framework for discussing probability problems, using the terms just mentioned.

Definition: A **random experiment** is a mechanism that produces a definite outcome that cannot be predicted with certainty.
The **sample space** associated with a random experiment is the set of all possible outcomes. An event is a subset of the sample space.

Definition: Element and Occurrence
An event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E
Definition: (**probability**) The probability of an outcome e in a sample space S is a P between 0 and 1 that measures the likelihood that e will occur on a single trial of the corresponding random experiment. The value P=0 corresponds to the outcome being impossible and the value P=1 corresponds to the outcome e being certain.

Definition: probability of an event
The probability of an event A is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted P(A)
Conditional Probability: The **conditional probability** of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. This probability is written P(B|A), notation for the probability of B given A. In the case where events A and B are independent (where event A has no effect on the probability of event B), the conditional probability of event B given event A is simply the probability of event B, that is P(B).

If events A and B are not independent, then the probability of the intersection of A and B (the probability that both events occur) is defined by
P(A and B) = P(A)P(B|A).

From this definition, the conditional probability *P(B|A)* is easily obtained by dividing by *P(A)*:

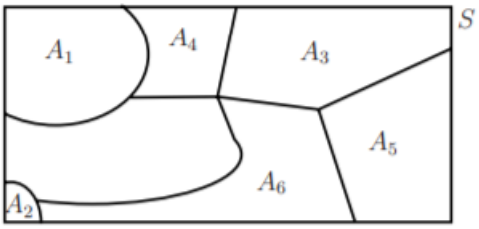
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Note: This expression is only valid when P(A) is greater than 0.

2.2 Total Probability and Bayes' Theorem
When the ideas of probability are applied to engineering (and many other areas) there are occasions when we need to calculate conditional probabilities other than those already known. For example, if production runs of ball bearings involve say, four machines, we might know the probability that any given machine produces faulty ball bearings. If we are inspecting the total output prior to distribution to users, we might need to know the probability that a faulty ball bearing came from a particular machine. Even though we do not address the area of statistics known as Bayesian Statistics here, it is worth noting that Bayes' theorem is the basis of this branch of the subject.
The theorem of total probability
To establish this result, we start with the definition of a partition of a sample space.
A partition of a sample space
The collection of events $A_1, A_2, \ldots A_n$ is said to **partition** a sample space S if

(a) $A_1 \cup A_2 \cup \cdots \cup A_n = S$
(b) $A_i \cap A_j = \emptyset$ for all i, j
(c) $A_i \neq \emptyset$ for all i

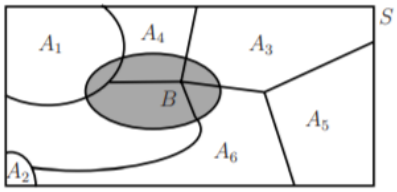
In essence, a partition is a collection of non-empty, non-overlapping subsets of a sample space whose union is the sample space itself. The definition is illustrated by Figure



If B is any event within S then we can express B as the union of subsets:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

The definition is illustrated in Figure 11 in which an event B in S is represented by the shaded region.



The bracketed events $(B \cap A_1), (B \cap A_2) \ldots (B \cap A_n)$ are mutually exclusive (if one occurs then none of the others can occur) and so, using the addition law of probability for mutually exclusive events:

$$\mathrm{P}(B) = \mathrm{P}(B \cap A_1) + \mathrm{P}(B \cap A_2) + \cdots + \mathrm{P}(B \cap A_n)$$

Each of the probabilities on the right-hand side may be expressed in terms of conditional probabilities:

$$\mathrm{P}(B \cap A_i) = \mathrm{P}(B|A_i)\mathrm{P}(A_i) \qquad \text{for all } i$$

Using these in the expression for $\mathrm{P}(B)$, above, gives:

$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(B|A_1)\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\mathrm{P}(A_2) + \cdots + \mathrm{P}(B|A_n)\mathrm{P}(A_n) \\ &= \sum_{i=1}^{n} \mathrm{P}(B|A_i)\mathrm{P}(A_i) \end{aligned}$$

This is the theorem of Total Probability. A related theorem with many applications in statistics can be deduced from this, known as Bayes' theorem.

Bayes' theorem

We again consider the conditional probability statement:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B|A_1)\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\mathrm{P}(A_2) + \cdots + \mathrm{P}(B|A_n)\mathrm{P}(A_n)}$$

in which we have used the theorem of Total Probability to replace $\mathrm{P}(B)$. Now

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

Substituting this in the expression for $\mathrm{P}(A|B)$ we immediately obtain the result

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B|A_1)\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\mathrm{P}(A_2) + \cdots + \mathrm{P}(B|A_n)\mathrm{P}(A_n)}$$

This is true for *any* event A and so, replacing A by A_i gives the result, known as Bayes' theorem as

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i) \times \mathrm{P}(A_i)}{\mathrm{P}(B|A_1)\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\mathrm{P}(A_2) + \cdots + \mathrm{P}(B|A_n)\mathrm{P}(A_n)}$$

Example:1

At a certain university, 4% of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman?

Solution

Let M={Student is Male}, F={Student is Female}.
Note that M and F partition the sample space of students.
Let T={Student is over 6 feet tall}.

We know that $\mathsf{P}(M) = 2/5$, $\mathsf{P}(F) = 3/5$, $\mathsf{P}(T|M) = 4/100$ and $\mathsf{P}(T|F) = 1/100$.

We require $\mathsf{P}(F|T)$. Using Bayes' theorem we have:

$$\begin{aligned}\mathsf{P}(F|T) &= \frac{\mathsf{P}(T|F)\mathsf{P}(F)}{\mathsf{P}(T|F)\mathsf{P}(F) + \mathsf{P}(T|M)\mathsf{P}(M)} \\ &= \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} \\ &= \frac{3}{11}\end{aligned}$$

Example:2

A factory production line is manufacturing bolts using three machines, A, B and C. Of the total output, machine A is responsible for 25%, machine B for 35% and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B and 2% from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from (a) machine A (b) machine B (c) machine C?

Solution

Let

D={bolt is defective},

A={bolt is from machine A},

B={bolt is from machine B},

C={bolt is from machine C}.

We know that $\mathsf{P}(A) = 0.25$, $\mathsf{P}(B) = 0.35$ and $\mathsf{P}(C) = 0.4$.

Also

$$\mathsf{P}(D|A) = 0.05, \mathsf{P}(D|B) = 0.04, \mathsf{P}(D|C) = 0.02.$$

A statement of Bayes' theorem for three events A, B and C is

$$\begin{aligned}\mathsf{P}(A|D) &= \frac{\mathsf{P}(D|A)\mathsf{P}(A)}{\mathsf{P}(D|A)\mathsf{P}(A) + \mathsf{P}(D|B)\mathsf{P}(B) + \mathsf{P}(D|C)\mathsf{P}(C)} \\ &= \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}\end{aligned}$$

$$= 0.362$$

Similarly

$$P(B|D) = \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= 0.406$$

$$P(C|D) = \frac{0.02 \times 0.4}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= 0.232$$

2.3 RANDOM VARIABLE:

A random variable is a number generated by a random experiment. A random variable is called discrete if its possible values form a finite or countable set. A random variable is called continuous if its possible values contain a whole interval of numbers

Example:3

In the experiment of throwing a coin twice the sample space S is S = {HH, HT, TH, TT}.

Let X be a random variable chosen such that X(S) = x (the number of heads).

Notation

Let 'S' be a sample space. The set of all outcomes 'S' in S such that X(S) = x is denoted by writing X = x.

$P(X = x) = P\{S : X(s) = x\}$

$P(x \leq a) = P\{S : X(s) \in (-\infty, a)\}$ and $P(a < x \leq b) = P\{s : X(s) \in (a, b)\}$

$P(X = a \text{ or } X = b) = P\{(X = a) \cup (X = b)\}$

$P(X = a \text{ and } X = b) = P\{(X = a) \cap (X = b)\}$ and so on.

Remark

The meaning of $P(X \leq a)$.

$P(X \leq a)$ is simply the probability of the set of outcomes 'S' in the sample space for which $X(s) \leq a$. Or $P(X \leq a) = P\{S : X(S) \leq a\}$

In the above example : we should write

$P(X \leq 1) = P(HH, HT, TH) = ¾$

Here $P(X \leq 1) = ¾$ means the probability of the R.V.X (the number of heads) is less than or equal to 1 is ¾ .

2.4 DISCRETE RANDOM VARIABLE

A random variable which can take only finite number of values or countably infinite number of values is called **discrete random** variable.

PROBABILITY MASS FUNCTION (PMF):

Consider a discrete random variable X which takes values x_1 , x_2 , x_3 ,……. It to each value x_i , we associate a number $p_i = P(X = x_i)$ then p_i is called the Probability Mass Function or Probability Function of random variable X, provided $p_i, i = 1,2, \ldots \ldots$ satisfies the following conditions i) $p_i \geq 0, \forall\, i$ and ii) $\sum_{i=1}^{\infty} p_i = 1$.

The function p(x) satisfying the above two conditions is called the probability mass function (or) probability distribution of the R.V.X. The probability distribution {xi, pi} can be displayed in the form of table as shown below.

$X = x_i$	x_1	x_2	…….	x_i
$P(X = x_i) = p_i$	p_1	p_2	…….	p_i

Example:4

If a random variable X takes the values 1,2,3,4 such that P(X=1)=3P(X=2)=P(X=3)=5P(X=4). Find the probability distribution of X.

Solution:

Assume P(X=3) = α. By the given equation, $P(X=1)=\frac{\alpha}{2}$ $P(X=2)=\frac{\alpha}{3}$ $P(X=4)=\frac{\alpha}{5}$.

For a probability distribution (and mass function) $\sum P(x) = 1$

P(1)+P(2)+P(3)+P(4) =1

$$\frac{\alpha}{2}+\frac{\alpha}{3}+\alpha+\frac{\alpha}{5}=1 \quad \Rightarrow \quad \frac{61}{30}\alpha=1 \quad \Rightarrow \alpha=\frac{30}{61}$$

$$P(X=1)=\frac{15}{61}\ ; P(X=2)=\frac{10}{61};\ P(X=3)=\frac{30}{61};\ P(X=4)=\frac{6}{61}$$

The probability distribution is given by

X	1	2	3	4
$p(x)$	$\frac{15}{61}$	$\frac{10}{61}$	$\frac{30}{61}$	$\frac{6}{61}$

Example:5

A random variable X has the following probability distribution.

X:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
f(x):	**0**	***k***	***2k***	***2k***	***3k***	k^2	$2k^2$	$7k^2+k$

Find (i) the value of *k* (ii) p(1.5 < X < 4.5 | X > 2)

Solution:

$$(i) \quad \sum P(x) = 1$$

$$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$10k^2 + 9k - 1 = 0$$

$$\Rightarrow k = -1, \frac{1}{10}$$

$$k = \frac{1}{10} = 0.1$$

$$(ii)\, P(1.5 < X < 4.5 \mid X > 2) = \frac{P(X=3) + P(X=4)}{P(X=3) + P(X=4) + P(X=5) + P(X=6) + P(X=7)}$$

$$= \frac{2k + 3k}{2k + 3k + k^2 + 2k^2 + 7k^2 + k}$$

$$= \frac{5k}{10k^2 + 6k} = \frac{\frac{5}{10}}{\frac{7}{10}} = \frac{5}{7}$$

2.5 CONTINUOUS DISTRIBUTION FUNCTION

A random variable which can take unaccountably infinite values or all values in an interval is called a continuous random variable.

PROBABILITY DENSITY FUNCTION (PDF) :

For a continuous random variable X, a probability density function is a function such that i) $f(x) \geq 0$ ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Remark

If f(x) is p.d.f of the random variable X then the probability that a value of the random variable X will fall in some interval (a, b) is equal to the definite integral of the function f(x) a to b.

$$P(a < x < b) = \int_a^b f(x)dx \qquad \text{(or)}$$

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

PROPERTIES OF P.D.F

The p.d.f f(x) of a random variable X has the following properties

- In the case of discrete random variable the probability at a point say at x = c is not zero. But in the case of a continuous random variable X the probability at a point is always zero.

$$P(X = c) = \int_{-\infty}^{\infty} f(x)dx = [x]_c^c = C - C = 0$$

- If x is a continuous random variable then we have $p(a \le X \le b) = p(a \le X < b) = p(a < X \le b)$

Example:6

Is the function defined as follows a density function?

$$f(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{18}(3+2x), & 2 \le x \le 4 \\ 0, & x > 4 \end{cases}$$

Solution:

$$\int_2^4 f(x)\ dx = \int_2^4 \frac{1}{18}(3+2x)\ dx = \left[\frac{(3+2x)^2}{72}\right]_2^4 = 1$$

Hence it is density function.

Example:7

A continuous random variable X has the probability function $f(x) = k(1+x),\ 2 \le x \le 5$

Find P(X<4).

Solution:

$$\int_2^4 f(x)dx = 1 \quad \Rightarrow k\int_2^5 (1+x)dx = 1 \Rightarrow k\left[\frac{(1+x)^2}{2}\right]_2^5 = 1 \Rightarrow k\frac{27}{2} = 1 \Rightarrow k = \frac{2}{27}$$

$$P(X < 4) = \int_2^4 f(x)\ dx = \frac{2}{27}\int_2^4 (1+x)\ dx = \frac{2}{27}\left[\frac{(1+x)^2}{2}\right]_2^4 = \frac{1}{25}(25-9) = \frac{16}{27}$$

Example:8

If the density function of a continuous random variable X is given by

$$f(x) = \begin{cases} ax & ;\ 0 \le x \le 1 \\ a & ;\ 1 \le x \le 2 \\ 3a - ax; & 2 \le x \le 3 \\ 0 & ;otherwise \end{cases}$$ then find the value of 'a'

$$\int_0^1 ax\,dx + \int_1^2 a\,dx + \int_2^3 (3a - ax)dx = 1 \qquad \Rightarrow a = \frac{1}{2}$$

Example:9

Given the p.d.f of a continuous R.V X as follows $f(x) = \begin{cases} 12.5x - 1.25 & 0.1 \le x \le 0.5 \\ 0, & elsewhere \end{cases}$

Find P(0.2 < X < 0.3)

Solution:

$$P(0.2 < X < 0.3) = \int_{0.2}^{0.3} (12.5x - 1.25)\, dx = \left[12.5\frac{x^2}{2} - 1.25x \right]_{0.2}^{0.3}$$
$$= 1.25\left[5(0.3)^2 - 0.3 - 5(0.2)^2 + 0.2\right]$$
$$= 0.1875$$

2.6 Cumulative Distribution function

The Cumulative Distribution Function (cdf) or Distribution Function of a random variable X (discrete or continuous) is given by

i) For a discrete random variable X: $F(x) = \sum_j P(X = x_j)\ \forall\, x_j \le x$

ii) For a continuous random variable X: $F(x) = P(-\infty < X \le x) = \int_{-\infty}^{x} f(x)dx$

Note

Let the random variable X takes values x1, x2, ….., xn with probabilities P1, P2, ….., Pn and let x1< x2< ….. <xn

Then we have

F(x) = P(X < x_1) = P(X < x1) + P(X = x1) = 0 + p_1 = p_1

F(x) = P(X < x2) = P(X < x1) + P(X = x1) + P(X = x2) = p_1 + p_2

F(x) = P(X < xn) = P(X < x_1) + P(X = x_2) + ….. + P(X = xn)

= p_1 + p_2+ ……. + p_n = 1

PROPERTIES OF DISTRIBUTION FUNCTIONS

Property: 1 P(a < X ≤ b) = F(b) – F(a), where F(x) = P(X ≤ x)

Property: 2 P(a ≤ X ≤ b) = P(X = a) + F(b) – F(a)

Property: 3 P(a < X < b) = P(a < X ≤ b) - P(X = b)

= F(b) – F(a) – P(X = b) by prob (1)

Example:10

Let X be a continuous random variable having the probability density function $f(x) = \begin{cases} \frac{2}{x^3}, & x \ge 1 \\ 0, & otherwise \end{cases}$ Find the distribution function of x.

Solution:

$$F(x) = \int_1^x f(x)\, dx = \int_1^x \frac{2}{x^3} dx = \left[-\frac{1}{x^2} \right]_1^x = 1 - \frac{1}{x^2}$$

Example:11

A random variable X has the probability density function *f(x)* given by $f(x)=\begin{cases} cxe^{-x}, & x\geq 0 \\ 0, & otherwise \end{cases}$. Find the value of c and CDF of X.

Solution:

$$\int_0^\infty f(x)\,dx=1\Rightarrow\int_0^\infty cxe^{-x}\,dx=1\Rightarrow c\left[-xe^{-x}-e^{-x}\right]_0^\infty=1\Rightarrow c(1)=1\Rightarrow c=1$$

$$F(x)=\int_0^x f(x)\,dx=\int_0^x cxe^{-x}\,dx=\int_0^x xe^{-x}\,dx=\left[-xe^{-x}-e^{-x}\right]_0^x=1-xe^{-x}-e^{-x}$$

Example:12

A continuous random variable X has the probability density function *f(x)* given by $f(x)=ce^{-|x|},-\infty<x<\infty$. Find the value of c and CDF of X.

Solution:

$$\int_{-\infty}^\infty f(x)\,dx=1\Rightarrow\int_{-\infty}^\infty ce^{-|x|}\,dx=1\Rightarrow 2\int_0^\infty ce^{-|x|}\,dx=1$$

$$\Rightarrow 2\int_0^\infty ce^{-x}\,dx=1\Rightarrow 2c\left[-e^{-x}\right]_0^\infty=1\Rightarrow 2c(1)=1\Rightarrow c=\frac{1}{2}$$

Case (i) $x<0$

$$F(x)=\int_{-\infty}^x f(x)\,dx=\int_{-\infty}^x ce^{-|x|}\,dx=c\int_{-\infty}^x e^{x}\,dx=c\left[e^{x}\right]_{-\infty}^x=\frac{1}{2}e^{x}$$

Case (ii) $x>0$

$$F(x)=\int_{-\infty}^x f(x)\,dx=\int_{-\infty}^x ce^{-|x|}\,dx=c\int_{-\infty}^0 e^{x}\,dx+c\int_0^x e^{-x}dx=c\left[e^{x}\right]_{-\infty}^0+c\left[-e^{-x}\right]_0^x$$

$$=c-ce^{-x}+c=c\left(2-e^{-x}\right)=\frac{1}{2}\left(2-e^{-x}\right)$$

Case (ii) $x>0$

$$F(x)=\int_{-\infty}^x f(x)\,dx=\int_{-\infty}^x ce^{-|x|}\,dx=c\int_{-\infty}^0 e^{x}\,dx+c\int_0^x e^{-x}dx=c\left[e^{x}\right]_{-\infty}^0+c\left[-e^{-x}\right]_0^x$$

$$=c-ce^{-x}+c=c\left(2-e^{-x}\right)=\frac{1}{2}\left(2-e^{-x}\right)$$

$$F(x)=\begin{cases}\frac{1}{2}e^{x}, & x>0\\ \frac{1}{2}\left(2-e^{-x}\right), & x<0\end{cases}$$

2.7 Mathematical Expectations:

The mathematical expectation or expected value or the mean value of X is defined as

$$\bar{X} = E(X) = \sum x_i p_i \quad \text{(Discrete random variable)}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad \text{(Continuous random variable)}$$

VARIANCE:

The mathematical expectation of $(X - \bar{X})^2$ is called as the variance of the distribution.

Thus, $Var(X) = E(X^2) - [E(X)]^2$.

Example:13

The number of hardware failures of a computer system in a week of operations has the following probability mass function:

No of failures	:	0	1	2	3	4	5	6
Probability	:	0.18	0.28	0.25	0.18	0.06	0.04	0.01

Find the mean of the number of failures in a week.

Solution:

$$E(X) = \sum x\ P(x) = (0)(0.18) + (1)(0.28) + (2)(0.25) + (3)(0.18) + (4)(0.06) + (5)(0.04) + (6)(0.01)$$
$$= 1.92$$

Example:14

The monthly demand for Titan watches is known to have the following probability distribution.

Demand:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Probability:	**0.08**	**0.3*k***	**0.19**	**0.24**	k^2	**0.1**	**0.07**	**0.04**

Determine the expected demand for watches. Also compute the variance.

Solution:

$$\sum P(x) = 1$$

$$(0.08) + (0.3k) + (0.19) + (0.24) + (k^2) + (0.1) + (0.07) +)(0.04) = 1$$

$$k^2 + 0.3k - 0.28 = 0$$

$$\Rightarrow k = 0.4$$

$$E(X) = \sum x\ P(x)$$

$$= (1)(0.18) + (2)(0.12) + (3)(0.19) + (4)(0.24) + (5)(0.16) + (6)(0.1) + (7)(0.07) + (8)(0.04)$$

Mean = $E(X) = 4.02$

$$E(X^2) = \sum x^2 P(x) = (1)(0.18) + (4)(0.12) + (9)(0.19) + (16)(0.24) + (25)(0.16) + (36)(0.1) + (49)(0.07) + (64)(0.04) = 19.7$$

$$Variance = E(X^2) - E(X)^2 = 19.07 - 4.02^2 = 3.54$$

Example :15

When die is thrown, 'X' denotes the number that turns up. Find E(X), E(X^2) and Var (X).

Solution Let 'X' be the random variable denoting the number that turns up in a die. 'X' takes values 1, 2, 3, 4, 5, 6 and with probability 1/6 for each

X = x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6
	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	$p(x_5)$	$p(x_6)$

Now

$$E(X) = \sum_{i=1}^{6} x_i\, p(x_i)$$

$$= x_1p(x_1) + x_2p(x_2) + x_3p(x_3) + x_4p(x_4) + x_5p(x_5) + x_6p(x_6)$$

$$= 1 \times (1/6) + 1 \times (1/6) + 3 \times (1/6) + 4 \times (1/6) + 5 \times (1/6) + 6 \times (1/6)$$

$$= 21/6 = 7/2 \qquad (1)$$

$$E(X) = \sum_{i=1}^{6} x_i\, p(x_p)$$

$$= x_1^2p(x_1)+x_2^2p(x_2)+x_3^2p(x_3)+x_4^2p(x_4)+x_5^2p(x_5)+x_6p(x_6)$$

$$= 1(1/6) + 4(1/6) + 9(1/6) + 16(1/6) + 25(1/6) + 36(1/6)$$

$$= \frac{1+4+9+16+25+36}{6} = \frac{91}{6} \qquad (2)$$

Variance (X) = Var (X) = $E(X^2) - [E(X)]^2$

$$= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

Example:16

A random variable X has the following probability function

Values of X	0	1	2	3	4	5	6	7	8
Probability P(X)	a	3a	5a	7a	9a	11a	13a	15a	17a

(i) Determine the value of 'a'

(ii) Find P(X<3), P(X≥3), P(0<X<5)

(iii) Find the distribution function of X.

Solution

(i) We know that if p(x) is the probability of mass function then

$$\sum_{i=0}^{8} p(x_i) = 1$$

$$p(0) + p(1) + p(2) + p(3) + p(4) + p(5) + p(6) + p(7) + p(8) = 1$$

$$a + 3a + 5a + 7a + 9a + 11a + 13a + 15a + 17a = 1$$

$$81\,a = 1$$

$$a = 1/81$$

Put a=1/81 , we get

X = x	0	1	2	3	4	5	6	7	8
P(x)	1/81	3/81	5/81	7/81	9/81	11/81	13/81	15/81	17/81

(ii) $P(X < 3) = p(0) + p(1) + p(2)$
$= 1/81 + 3/81 + 5/81 = 9/81$
(ii) $P(X \geq 3) = 1 - p(X < 3)$
$= 1 - 9/81 = 72/81$
(iii) $P(0 < x < 5) = p(1) + p(2) + p(3) + p(4)$ here 0 & 5 are not include
$= 3/81 + 5/81 + 7/81 + 9/81$

$$= \frac{3 + 5 + 7 + 8 + 9}{81} = \frac{24}{81}$$

(iv) To find the distribution function of X using table 2, we get

X = x	**F(X) = P(x ≤ x)**
0	F(0) = p(0) = 1/81
1	F(1) = P(X ≤ 1) = p(0) + p(1) = 1/81 + 3/81 = 4/81
2	F(2) = P(X ≤ 2) = p(0) + p(1) + p(2) = 4/81 + 5/81 = 9/81
3	F(3) = P(X ≤ 3) = p(0) + p(1) + p(2) + p(3) = 9/81 + 7/81 = 16/81
4	F(4) = P(X ≤ 4) = p(0) + p(1) + …. + p(4) = 16/81 + 9/81 = 25/81
5	F(5) = P(X ≤ 5) = p(0) + p(1) + ….. + p(4) + p(5) = 2/81 + 11/81 = 36/81
6	F(6) = P(X ≤ 6) = p(0) + p(1) + ….. + p(6) = 36/81 + 13/81 = 49/81
7	F(7) = P(X ≤ 7) = p(0) + p(1) + …. + p(6) + p(7) = 49/81 + 15/81 = 64/81
8	F(8) = P(X ≤ 8) = p(0) + p(1) + ….. + p(6) + p(7) + p(8) = 64/81 + 17/81 = 81/81 = 1

Example:17

If X has the probability density function $f(x) = k e^{-3x},\ x > 0$

Find (i) k (ii) p(0.5 ≤ X ≤ 1) (iii) Mean of X.

Solution:

(i) $\int_0^\infty f(x)\,dx = 1$

$$\int_0^\infty k e^{-3x}\,dx = 1$$

$$k\left[\frac{-e^{-x}}{3}\right]_0^\infty = 1$$

$$k\left(\frac{1}{3}\right) = 1$$

$$k = 3$$

(ii) $p(0.5 \le X \le 1) = \int_{0.5}^{1} f(x)\,dx = \int_{0.5}^{1} 3e^{-3x}\,dx$

$$= 3\left[\frac{-e^{-3x}}{3}\right]_{0.5}^{1} = -e^{-3} + e^{-1.5}$$

(iii) $Mean = E(X) = \int_0^\infty x f(x)\,dx = \int_0^\infty 3x e^{-3x}\,dx$

$$= 3\left[-x\frac{e^{-3x}}{3} - \frac{e^{-3x}}{9}\right]_0^\infty = 3\left(\frac{1}{9}\right) = \frac{1}{3}$$

Example:18

If X has the distribution function

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{3}, & 1 \le x < 4 \\ \frac{1}{2}, & 4 \le x < 6 \\ \frac{5}{6}, & 6 \le x < 10 \\ 1, & x > 10 \end{cases}$$

Find (1) Probability distribution of X (2) p(2 < X < 6) (3) Mean (4) variance

Solution:

(1) As there are no x terms in the distribution function given is a discrete random variable.

Hence the probability distribution is given by

X	1	4	6	10
$p(X)$	$\frac{1}{3}$	$\frac{1}{2}-\frac{1}{3}$	$\frac{5}{6}-\frac{1}{2}$	$1-\frac{5}{6}$
	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

(2) p(2 < X < 6) = p(4) = $\frac{1}{6}$

(3) Mean = E(X) = $\sum x\ p(x) = (1)\left(\frac{1}{3}\right)+(4)\left(\frac{1}{6}\right)+(6)\left(\frac{1}{3}\right)+(10)\left(\frac{1}{6}\right) = \frac{14}{3}$

(4) $E(X^2) = \sum x^2\ p(x) = (1)\left(\frac{1}{3}\right)+(16)\left(\frac{1}{6}\right)+(36)\left(\frac{1}{3}\right)+(100)\left(\frac{1}{6}\right) = \frac{95}{3}$

Variance = $E(X^2) - [E(X)]^2 = \frac{95}{3} - \frac{196}{9} = \frac{89}{9}$

2.8 Moments and Moment Generating Function

Moments about Origin (Or Raw Moment):

r^{th} moment about origin i.e., $\mu_r' = E(X^r)$

i.e., $\mu_r' = \sum X^r P(X)$, if X is discrete

i.e., $\mu_r' = \int_{-\infty}^{\infty} X^r f(x)dx$, if X is continuous

Note:

(i) $\mu_0' = \int_0^{\infty} f(x)\,dx = 1$

(ii) $\mu_1' = \int_0^{\infty} xf(x)\,dx = E(X) = Mean$

(iii) $\mu_2' = \int_0^{\infty} x^2 f(x)\,dx = E(X^2)$

CENTRAL MOMENTS:

For a r.v X, discrete and continuous, its r^{th} moment about any number A is defined as, r^{th} moment about $A = E(X - A)^r$.

i.e., $\mu_r = \sum_x (x - \bar{X})^r P(x)$, if X is discrete

i.e., $\mu_r = \int_{-\infty}^{\infty} (x - \bar{X})^r f(x) dx$, if X is continuous

Note:

(i) $\mu_0 = \int_{-\infty}^{\infty} f(x)\, dx = 1$

(ii) $\mu_1 = \int_{-\infty}^{\infty} (x - \overline{X}) f(x)\, dx = 0$

(iii) $\mu_2 = \int_{0}^{\infty} (x - \overline{X})^2 f(x)\, dx = Var(X)$

(iv) $\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_1'^3$

(v) $\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_1'^2 \mu_2' - 3\mu_1'^4$

Example:19

The density function of a random variable X is given by

$$f(x) = kx(2 - x), \quad 0 \le x \le 2$$

Find *k*, mean, variance and *r*th moment.

Solution:

$$\int_0^2 f(x)\, dx = 1$$

$$\int_0^2 kx(2 - x)\, dx = 1$$

$$k\int_0^2 (2x - x^2)\, dx = 1 \qquad \Rightarrow k\left[x^2 - \frac{x^3}{3}\right]_0^2 = 1$$

$$k\left(4 - \frac{8}{3}\right) = 1 \quad \Rightarrow k = \frac{3}{4}$$

$$\mu_r' = \int_0^2 x^r \frac{3}{4}x(2-x)\,dx$$

$$= \frac{3}{4}\int_0^2 \left(2x^{r+1} - x^{r+2}\right)dx$$

$$= \frac{3}{4}\left[2\frac{x^{r+2}}{r+2} - \frac{x^{r+3}}{r+3}\right]_0^2 \qquad = \frac{3}{4}\left[\frac{2^{r+3}}{r+2} - \frac{2^{r+3}}{r+3}\right]$$

$$= \frac{3}{4}2^{r+3}\left[\frac{1}{r+2} - \frac{1}{r+3}\right]$$

$$= 6\left(2^r\right)\frac{1}{(r+2)(r+3)}$$

put r = 1,2 $\qquad \mu_1' = \frac{12}{(3)(4)} = 1 \qquad \mu_2' = \frac{24}{(4)(5)} = \frac{6}{5}$

Mean = 1 and variance = $\mu_{2'} - \mu_{1'^2} = \frac{6}{5} - 1 = \frac{1}{5}$

MOMENT GENERATING FUNCTION (MGF):

The moment generating function (m.g.f) of a random variable X (about origin), which is denoted as $M_X(t)$, is defined as $M_X(t) = E(e^{tX})$, where t is a real parameter $(-\infty < t < \infty)$, which is independent of X.

NOTE:

$M_X(t) = E\left(e^{tX}\right) = \sum_x e^{tX} P(x)$, X is discrete.

$= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$, X is continuous.

EXPECTATION TABLE

Discrete R.V's	Continuous R.V's
1. $E(X) = \sum x\, p(x)$	1. $E(X) = \int_{-\infty}^{\infty} x f(x) dx$
2. $E(X^r) = \mu_r' = \sum_x x^r p(x)$	2. $E(X^r) = \mu_r' = \int_{-\infty}^{\infty} x^r f(x) dx$
3. Mean $= \mu_r' = \sum x\, p(x)$	3. Mean $= \mu_r' = \int_{-\infty}^{\infty} x\, f(x) dx$
4. $\mu_2' = \sum x^2 p(x)$	4. $\mu_2' = \int_{-\infty}^{\infty} x^2 f(x) dx$
5. Variance $= \mu_2' - \mu_1'^2 = E(X^2) - \{E(X)\}^2$	5. Variance $= \mu_2' - \mu_1'^2 = E(X^2) - \{E(X)\}^2$

Example:20

Find the MGF of triangular distribution whose density function is given by

$$f(x) = \begin{cases} x, & 0 < x < 1 \\ 2 - x, & 1 < x < 2 \\ 0, & elsehwere \end{cases}$$

Hence its mean and variance.

Solution:

$$M_X(t) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

$$= \int_0^1 e^{tx}\, x\, dx + \int_1^2 e^{tx}(2 - x)\, dx$$

$$= \left[x\frac{e^{tx}}{t} - \frac{e^{tx}}{t^2}\right]_0^1 + \left[(2 - x)\frac{e^{tx}}{t} - (-1)\frac{e^{tx}}{t^2}\right]_1^2$$

$$= \frac{e^t}{t} - \frac{e^t}{t^2} + \frac{1}{t^2} + \frac{e^{2t}}{t^2} - \frac{e^t}{t} - \frac{e^t}{t^2}$$

$$M_X(t) = \frac{e^{2t} - 2e^t + 1}{t^2}$$

Expanding the above in powers of t, we get

$$M_X(t)=\frac{e^{2t}-2e^{t}+1}{t^2}=\frac{1}{t^2}\left[\begin{array}{l}\left(1+2t+\frac{4t^2}{2!}+\frac{8t^3}{3!}+\frac{16t^4}{4!}+\ldots\right)\\ -2\left(1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+.\right)-1\end{array}\right]$$

$$=\frac{1}{t^2}\left(\frac{2t^2}{2!}+\frac{6t^3}{3!}+\frac{14t^4}{4!}+\ldots\right)$$

$$=1+t+\frac{7t^2}{12}+\frac{t^3}{4}+\ldots.$$

Mean = E(X) = (coefficient of t) 1! = 1

$$E\left(X^2\right)=\text{(coefficient of } t^2\text{)}2!=\frac{7}{6}$$

Example:21

Find MGF of the RV X, whose p.d.f is given by

$$f(x)=\lambda e^{-\lambda x},\ x>0$$

hence find the first four central moments.

Solution:

$$M_X(t)=E\left(e^{tX}\right)\quad=\int_{-\infty}^{\infty}e^{tx}f(x)\,dx$$

$$=\int_0^{\infty}e^{tx}\lambda e^{-\lambda x}\,dx\ =\lambda\int_0^{\infty}e^{-(\lambda-t)x}\,dx$$

$$=\lambda\left[\frac{e^{-(\lambda-t)x}}{-(\lambda-t)}\right]_0^{\infty}=\frac{\lambda}{(\lambda-t)}$$

Expanding in powers of t

$$M_X(t)=\frac{\lambda}{(\lambda-t)}=\frac{1}{1-\left(\frac{t}{\lambda}\right)}=1+\left(\frac{t}{\lambda}\right)+\left(\frac{t}{\lambda}\right)^2+\left(\frac{t}{\lambda}\right)^3+\ldots$$

taking the coefficient we get the raw moments about origin

$$E(X) = (coefficien\ t\ \ of\ \ t)1! = \frac{1}{\lambda}$$

$$E(X^2) = (coefficien\ t\ of\ t^2)2! = \frac{2}{\lambda^2}$$

$$E(X^3) = (coefficien\ t\ \ of\ \ t^3)3! = \frac{6}{\lambda^3}$$

$$E(X^4) = (coefficien\ t\ \ of\ \ t^4)4! = \frac{24}{\lambda^4}$$

and the central moments are

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - 2C_1\mu_1'\mu_1' + \mu_1'^2$$

$$= \frac{2}{\lambda^2} - 2\frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$$\mu_3 = \mu_3' - 3C_1\mu_2'\mu_1' + 3C_2\mu_1'\mu_1'^2 - \mu_1'^3$$

$$= \frac{6}{\lambda^3} - 3\frac{2}{\lambda^2}\frac{1}{\lambda} + 3\frac{1}{\lambda}\frac{1}{\lambda^2} - \frac{1}{\lambda^3} = \frac{2}{\lambda^3}$$

$$\mu_4 = \mu_4' - 4C_1\mu_3'\mu_1' + 4C_2\mu_2'\mu_1'^2 - 4C_3\mu_1'^4 + \mu_1'^4$$

$$= \frac{24}{\lambda^4} - 4\frac{6}{\lambda^3}\frac{1}{\lambda} + 6\frac{2}{\lambda^2}\frac{1}{\lambda^2} - 4\frac{1}{\lambda^4} + \frac{1}{\lambda^4} = \frac{9}{\lambda^4}$$

2.9 Discrete Distributions

The important discrete distribution of a random variable 'X' are

1. Binomial Distribution
2. Poisson Distribution
3. Geometric Distribution

2.9.1 BINOMIAL DISTRIBUTION

Definition: A random variable X is said to follow binomial distribution if its probability law is given by P(x) = p(X = x successes) = $nC_x p^x q^{n-x}$ Where x = 0, 1, 2, ……., n, p+q = 1

Note Assumptions in Binomial distribution

i) There are only two possible outcomes for each trail (success or failure).
ii) The probability of a success is the same for each trail.
iii) There are 'n' trails, where 'n' is a constant.
iv) The 'n' trails are independent.

Moment Generating Function (MGF) of a binomial distribution about origin.

WKT $M_X(t) = \sum_{x=0}^{n} e^{tx} p(x)$

Let 'X' be a random variable which follows binomial distribution then MGF about origin is given by

$$E[e^{tX}] = M_X(t) = \sum_{x=0}^{n} e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx} nC_x p^x q^{n-x} \qquad \left[\because p(x) = nC_x p^x q^{n-x}\right]$$

$$= \sum_{x=0}^{n} (e^{tx}) p^x nC_x q^{n-x}$$

$$= \sum_{x=0}^{n} (pe^t)^x nC_x q^{n-x}$$

$$\therefore M_X(t) = (q + pe^t)^n$$

Find the mean and variance of binomial distribution.

Solution

$$M_X(t) = (q + pe^t)^n$$

$$\therefore M_X'(t) = n(q + pe^t)^{n-1}.pe^t$$

Put $t = 0$, we get

$$M_X'(0) = n(q + p)^{n-1}.p$$

$$\text{Mean} = E(X) = np \qquad [\because (q + p) = 1] \qquad \left[\text{Mean } M_X'(0)\right]$$

$$M_X''(t) = np\left[(q + pe^t)^{n-1}.e^t + e^t(n-1)(q + pe^t)^{n-2}.pe^t\right]$$

Put $t = 0$, we get

$$M_X''(t) = np\left[(q + p)^{n-1} + (n-1)(q + p)^{n-2}.p\right]$$

$$= np[1 + (n-1)p]$$

$$= np + n^2p^2 - np^2$$

$$= n^2p^2 + np(1-p)$$

$$M_X''(0) = n^2p^2 + npq \qquad [\because 1 - p = q]$$

$$M_X''(0) = E(X^2) = n^2p^2 + npq$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = n^2/p^2 + npq - n^2/p^2 = npq$$

$$\text{Var}(X) = npq$$

$$\text{S.D} = \sqrt{npq}$$

Additive property of Binomial Distribution

The sum of two binomial variants is not a binomial variate.

Let X and Y be two independent binomial variates with parameter (n_1, p_1) and (n_2, p_2) respectively.

Then

$$M_X(t) = \left(q_1 + p_1 e^t\right)^{n_1}, \qquad M_Y(t) = \left(q_2 + p_2 e^t\right)^{n_2}$$

$$\therefore\ M_{X+Y}(t) = M_X(t).M_Y(t) \quad [\because \text{X \& Y are independent R.V.'s}]$$

$$= \left(q_1 + p_1 e^t\right)^{n_1}.\left(q_2 + p_2 e^t\right)^{n_2}$$

RHS cannot be expressed in the form $\left(q + pe^t\right)^n$. Hence by uniqueness theorem of MGF X+Y is not a binomial variate. Hence in general, the sum of two binomial variates is not a binomial variate.

If $M_X(t) = \left(q+pe^t\right)^{n_1}$, $M_Y(t) = \left(q+pe^t\right)^{n_2}$, then

$$M_{X+Y}(t) = \left(q+pe^t\right)^{n_1+n_2}$$

Example22

1. Check whether the following data follow a binomial distribution or not. Mean = 3; variance = 4.

Solution

Given Mean $np = 3$ (1)

Variance $npr = 4$ (2)

$$\frac{(2)}{(1)} \Rightarrow \frac{np}{npq} = \frac{3}{4}$$

$$\Rightarrow q = \frac{4}{3} = 1\frac{1}{3} \text{ which is} > 1.$$

Since $q > 1$ which is not possible $(0 < q < 1)$. The given data not follow binomial distribution.

Example:23

The mean and SD of a binomial distribution are 5 and 2, determine the distribution

Solution

Given Mean $= np = 5$ (1)

$SD = \sqrt{npq} = 2$ (2)

$$\frac{(2)}{(1)} \Rightarrow \frac{np}{npq} = \frac{4}{5} \Rightarrow q = \frac{4}{5}$$

$$\therefore p = 1 - \frac{4}{5} = \frac{1}{5} \Rightarrow p = \frac{1}{5}$$

Sub (3) in (1) we get

n x 1/5 = 5

n = 25

∴ The binomial distribution is

$$P(X = x) = p(x) = {}^{n}C_x p^x q^{n-x}$$

$$= 25C_x (1/5)^x (4/5)^{n-x}, \qquad x = 0, 1, 2, \ldots, 25$$

Example:24

It has been claimed that in 60 % of all solar heat installation the utility bill is reduced by atleast one-third. Accordingly what are the probabilities that the utility bill will be reduced by atleast one-third in atleast four of five installations?

Solution:

Given n=5, p=60 % =0.6 and q=1-p=0.4

$$\begin{aligned} p(x \geq 4) &= p[x=4] + p[x=5] \\ &= 5c_4 (0.6)^4 (0.4)^{5-4} + 5c_5 (0.6)^5 (0.4)^{5-5} \\ &= 0.337 \end{aligned}$$

Example:25

In a company 5 % defective components are produced. What is the probability that atleast 5 components are to be examined in order to get 3 defectives?

Solution:

To get 3 defectives, 3 or more components must be examined.

p=5 % =0.05, q = 1- p=0.95 and k=success=3

$$p(X=x) = (x-1)c_{k-1}\, p^k\, q^{x-k} \quad , x = k, k+1, k+2, \ldots$$

$$\begin{aligned} p(x \geq 5) &= 1 - p(x<5) \\ &= 1 - [p(x=3) + p(x=4)] \\ &= 1 - \left[2c_2 (0.05)^3 (0.95)^0 + 3c_2 (0.05)^3 (0.95)^1\right] \end{aligned}$$

2.9.2 Poisson Distribution

Definition: A random variable X is said to follow if its probability law is given by

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

Poisson distribution is a limiting case of binomial distribution under the following conditions or assumptions.

1. The number of trails ‘n’ should e infinitely large i.e. $n \to \infty$.
2. The probability of successes ‘p’ for each trail is infinitely small.
3. $np = \lambda$, should be finite where λ is a constant.

Find MGF of Poisson Distribution and whose Mean & Variance.

$$M_X(t) = E(e^{tx})$$

$$= \sum_{x=0}^{\infty} e^{tx} p(x)$$

$$= \sum_{x=0}^{\infty} e^{tx} \left(\frac{\lambda^x e^{\lambda}}{x!} \right)$$

$$= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} \left[1 + \lambda e^t + \frac{(\lambda e^t)^2}{2!} + \ldots\ldots \right]$$

$$= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

$= 0, 1, 2, \ldots\ldots, 25$

Hence $M_X(t) = e^{\lambda(e^t - 1)}$

* To find Mean and Variance

WKT $M_X(t) = e^{\lambda(e^t - 1)}$

$\therefore$ $M_X'(t) = e^{\lambda(e^t - 1)} . e^t$

$M_X'(0) = e^{-\lambda} . \lambda$

$\mu_1' = E(X) = \sum_{x=0}^{\infty} x.p(x)$

$$= \sum_{x=0}^{\infty} x . \frac{e^{-\lambda} \lambda^x}{x!} \quad = \sum_{x=0}^{\infty} \frac{x . e^{-\lambda} \lambda \lambda^{x-1}}{x!}$$

$$= 0 + e^{-\lambda} . \lambda \sum_{x=1}^{\infty} \frac{x . \lambda^{x-1}}{x!}$$

$$= \lambda e^{-\lambda} . \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \ldots\ldots \right]$$

$$= \lambda e^{-\lambda} . e^{\lambda}$$

Mean $= \lambda$

$$\mu_2' = E[X^2] = \sum_{x=0}^{\infty} x^2.p(x) = \sum_{x=0}^{\infty} x^2.\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} \{x(x-1)+x\}.\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} \frac{x(x-1)e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{x.e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)(x-3)....1} + \lambda$$

$$= e^{-\lambda}\lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= e^{-\lambda}\lambda^2 \left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + ...\right] + \lambda$$

$$= \lambda^2 + \lambda$$

Variance $\mu_2 = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Variance = λ

Hence Mean = Variance = λ

Note : * sum of independent Poisson Vairates is also Poisson variate.

Example:26

If X is a Poisson variate such tht P(X=1) 3/10 and P(X=2)=1/5, find P(X=0) and P(X=3)

Solution

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

$$\therefore P(X=1) = e^{-\lambda}\lambda = \frac{3}{10} \quad \text{(Given)}$$

$$\lambda e^{-\lambda} = \frac{3}{10} \quad (1)$$

$$P(X=2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{1}{5} \quad \text{(Given)}$$

$$\frac{e^{-\lambda}\lambda^2}{2!} = \frac{1}{5} \quad (2)$$

$$(1) \Rightarrow e^{-\lambda}\lambda = \frac{3}{10} \quad (3)$$

$$(2) \Rightarrow e^{-\lambda}\lambda^2 = \frac{2}{5} \quad (4)$$

$$\frac{(3)}{(4)} \Rightarrow \frac{1}{\lambda} = \frac{3}{4}$$

$$\lambda = \frac{4}{3}$$

$$\therefore P(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-4/3}$$

$$P(X=3) = \frac{e^{-\lambda}\lambda^3}{3!} = \frac{e^{-4/3}(4/3)^3}{3!}$$

Example:27

If X is a Poisson variable such that P(X=2)=9P(X=4)+90P(X=6) Find (i) Mean of X (ii) Variance of X

Solution

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

Given $P(X=2) = 9\,P(X=4) + 90\,P(X=6)$

$$\frac{e^{-\lambda}\lambda^{2}}{2!} = 9\frac{e^{-\lambda}\lambda^{4}}{4!} + 90\frac{e^{-\lambda}\lambda^{6}}{6!}$$

$$\frac{1}{2} = \frac{9\lambda^{2}}{4!} + \frac{90\lambda^{4}}{6!} \qquad (2)$$

$$\frac{1}{2} = \frac{3\lambda^{2}}{8} + \frac{\lambda^{4}}{8} \qquad (3)$$

$$1 = \frac{3\lambda^{2}}{4} + \frac{\lambda^{4}}{4} \qquad (4)$$

$$\frac{(3)}{(4)} \Rightarrow \frac{1}{\lambda} = \frac{3}{4}$$

$$\lambda = \frac{4}{3}$$

$$\therefore \quad P(X=0) = \frac{e^{-\lambda}\lambda^{0}}{0!} = e^{-4/3}$$

$$P(X=3) = \frac{e^{-\lambda}\lambda^{3}}{3!} = \frac{e^{-4/3}(4/3)^{3}}{3!}$$

$$\lambda^{4} + 3\lambda^{2} - 4 = 0$$

$\lambda^{2} = 1$ or $\lambda^{2} = -4$

$\lambda = \pm 1$ or $\lambda = \pm 2i$

$\therefore$ Mean = λ = 1, Variance = λ = 1

$\therefore$ Standard Deviation = 1

Example:28

It is known that the probability of an item produced by a certain machine will be defective is 0.05. If the produced items are sent to the market in packets of 20, find the no. of packets containing at least , exactly, at most 2 defectives in a consignment of 1000 packets using Poisson distribution.

Solution:

Give n = 20 , p = 0.05 , N = 1000

Mean $\lambda = np = 1$

Let X denotes the number of defectives.

$$p[X = x] = \frac{e^{-\lambda}.\lambda^x}{x!} = \frac{e^{-1}.1^x}{x!} = \frac{e^{-1}.}{x!} \quad x = 0,1,2,...$$

$$p[x \geq 2] = 1 - p[x < 2]$$
$$= 1 - [p(x = 0) + p(x = 1)]$$
$$= 1 - \left[\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!}\right] = 1 - 2e^{-1} = 0.2642$$

Therefore, out of 1000 packets , the no. of packets containing atleast 2 defectives

$$= N.p[x \geq 2] = 1000 * 0.2642 \cong 264 \ packets$$

(ii) $p[x = 2] = \frac{e^{-1}}{2!} = 0.18395$

Out of 1000 packets,=*N*P[x=2]*=184 packets.

(ii)

$$p[x \leq 2] = p[x = 0] + p[x = 1] + p[x = 2]$$
$$= \frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} = 0.91975$$

Example:29

The atoms of radioactive element are randomly disintegrating. If every gram of this element , on average, emits 3.9 alpha particles per second, what is the probability during the next second the no. of alpha particles emitted from 1 gram is (i) atmost 6 (ii) atleast 2 (iii) atleast 3 and atmost 6 ?

Solution:

Given $\lambda = 3.9$

Let X denote the no. of alpha particles emitted

$$(i)\ p(x \le 6) = p(x=0) + p(x=1) + p(x=2) + ... + p(x=6)$$
$$= \frac{e^{-3.9}(3.9)^0}{0!} + \frac{e^{-3.9}(3.9)^1}{1!} + \frac{e^{-3.9}(3.9)^2}{2!} + ... + \frac{e^{-3.9}(3.9)^6}{6!}$$
$$= 0.898$$

$$(ii)\quad p(x \ge 2) = 1 - p(x < 2)$$
$$= 1 - [p(x=0) + p(x=1)]$$
$$= 1 - \left[\frac{e^{-3.9}(3.9)^0}{0!} + \frac{e^{-3.9}(3.9)^1}{1!}\right]$$
$$= 0.901$$

$$(iii)\quad p(3 \le x \le 6) = p(x=3) + p(x=4) + p(x=5) + p(x=6)$$
$$= \frac{e^{-3.9}(3.9)^3}{3!} + \frac{e^{-3.9}(3.9)^4}{4!} + \frac{e^{-3.9}(3.9)^5}{5!} + \frac{e^{-3.9}(3.9)^6}{6!}$$
$$= 0.645$$

Example:30

The number of monthly breakdown of a computer is a random variable having Poisson `distribution with mean 1.8. Find the probability that this computer will function for a `month with only one breakdown.

Solution:

$$p(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}\ ,\quad given\ \lambda = 1.8,\ p(x=1) = \frac{e^{-1.8}(1.8)^1}{1!} = 0.2975$$

Example:31

A discrete R.V X has mgf $M_x(t) = e^{2(e^t-1)}$. Find E(x), var(x), and p(x=0).

Solution: Given $M_x(t) = e^{2(e^t-1)}$

We know that mgf of poisson is $M_x(t) = e^{\lambda(e^t-1)}$

Therefore $\lambda = 2$. In poisson E(x) = var(x) = λ

$$\therefore\ Mean\ E(x) = \text{var}(x) = 2$$

$$p(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\therefore\ p(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = e^{-2} = 0.1353$$

Derive probability mass function of Poisson distribution as a limiting case of Binomial distribution

Solution

We know that the Binomial distribution is $P(X=x) = nCx\ p^x q^{n-x}$

$$P(X=x) = nC_x\, p^x q^{n-x}$$

$$= \frac{n!}{(n-x)!\,x!} p^x (1-p)^{n-x}$$

$$= \frac{1.2.3.....(n-x)(n-x+1).....np^n}{1.2.3.....(n-x)\ x!} \frac{(1-p)^n}{(1-p)^x}$$

$$= \frac{1.2.3.....(n-x)(n-x+1).....n}{1.2.3.....(n-x)\ x!} \left(\frac{p}{1-p}\right)^x (1-p)^n$$

$$= \frac{n(n-1)(n-2).....(n-x+1)}{x!} \frac{\lambda^x}{n^x} \frac{1}{\left(1-\frac{\lambda}{n}\right)^x} \left(1-\frac{\lambda}{n}\right)^n$$

$$= \frac{n(n-1)(n-2).....(n-x+1)}{x!} x \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

$$P(X=x) = \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left\{1-\left(\frac{x-1}{n}\right)\right\}}{x!} \lambda^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left\{1-\left(\frac{x-1}{n}\right)\right\}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

When $n\to\infty$

$$P(X=x) = \frac{\lambda^x}{x!} \lim_{n\to\infty}\left[1-\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left\{1-\left(\frac{x-1}{n}\right)\right\}\left(1-\frac{\lambda}{n}\right)^{n-x}\right]$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty}\left(1-\frac{1}{n}\right) \lim_{n\to\infty}\left(1-\frac{2}{n}\right)..... \lim_{n\to\infty} 1-\left(\frac{x-1}{n}\right)$$

We know that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-x} = e^{-\lambda}$$

and $$\lim_{n\to\infty}\left(1-\frac{1}{n}\right) = \lim_{n\to\infty}\left(1-\frac{2}{n}\right)..... = \lim_{n\to\infty}\left(1-\left(\frac{x-1}{n}\right)\right) = 1$$

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

2.9.3 GEOMETRIC DISTRIBUTION

Definition: A discrete random variable 'X' is said to follow geometric distribution, if it assumes only non-negative values and its probability mass function is given by $P(X= x) = p(x) = q^{x-1}p$;x= 1,2,3,… 0<p<1, where q = 1-p

To find MGF

$M_X(t) = E[e^{tx}]$

$= \sum e^{tx} p(x)$

$= \sum_{x=1}^{\infty} e^{tx} q^{x-1} p$

$= \sum_{x=1}^{\infty} e^{tx} q^{x} q^{-1} p$

$= \sum_{x=1}^{\infty} e^{tx} q^{x} p / q$

$= p/q \sum_{x=1}^{\infty} e^{tx} q^{x}$

$= p/q \sum_{x=1}^{\infty} (e^{t}q)^{x}$

$= p/q \left[(e^{t}q)^{1} + (e^{t}q)^{2} + (e^{t}q)^{3} + ... \right]$

Let $x = e^{t}q$ $= p/q \left[x + x^{2} + x^{3} + ... \right]$

$= \frac{p}{q} x \left[1 + x + x^{2} + ... \right] = \frac{p}{q} (1-x)^{-1}$

$= \frac{p}{q} qe^{t} \left[1 - qe^{t} \right] = pe^{t} \left[1 - qe^{t} \right]^{-1}$

$\therefore M_X(t) = \frac{pe^{t}}{1 - qe^{t}}$

$$M_X'(t) = \frac{(1-qe^{t})pe^{t} - pe^{t}(-qe^{t})}{(1-qe^{t})^{2}} = \frac{pe^{t}}{(1-qe^{t})^{2}}$$

$\therefore E(X) = M_X'(0) = 1/p$

$\therefore$ Mean $= 1/p$

Variance $\mu_X''(t) = \frac{d}{dt}\left[\frac{pe^{t}}{(1-qe^{t})^{2}} \right]$

$$= \frac{(1-qe^{t})^{2} pe^{t} - pe^{t} 2(1-qe^{t})(-qe^{t})}{(1-qe^{t})^{4}}$$

$$= \frac{(1-qe^{t})^{2} pe^{t} + 2pe^{t} qe^{t} (1-qe^{t})}{(1-qe^{t})^{4}}$$

$$M_X''(0) = \frac{1+q}{p^{2}}$$

$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{(1+q)}{p^2} - \frac{1}{p^2} \Rightarrow \frac{q}{p^2}$

$\text{Var}(X) = \frac{q}{p^2}$

Note:

Another form of geometric distribution

$P[X=x] = q^{x}p$; $x = 0, 1, 2, ...$

$$M_X(t) = \frac{p}{(1-qe^{t})}$$

Mean $= q/p$, Variance $= q/p^2$

Establish the memory less property of geometric distribution.

$p[x > m+n/x > m] = p[x > n]$

Solution:

If X is a discrete r.v. following a geometric distbn.

$$\therefore p(X = x) = pq^{x-1} \quad , x = 1,2,.....$$

$$p(x > k) = \sum_{x=k+1}^{\infty} pq^{x-1}$$

$$= p\left[q^k + q^{k+1} + q^{k+2} + ...\right]$$

$$= p\,q^k\left[1 + q + q^2 + ...\right] = p\,q^k(1-q)^{-1}$$

$$= p\,q^k\,p^{-1} = q^k$$

Now,

$$p[x > m+n/x > m] = \frac{p[x > m+n \text{ and } x > m]}{p[x > m]}$$

$$= \frac{p[x > m+n]}{p[x > m]} = \frac{q^{m+n}}{q^m} = q^n = p[x > n]$$

$$\therefore p[x > m+n/x > m] = p[x > n]$$

Example:32

If the MGF of X is $(5-4et)^{-1}$, find the distribution of X and P(X=5)

Solution

Let the geometric distribution be

$P(X = x) = q^x p, \quad x = 0, 1, 2, ...$

The MGF of geometric distribution is given by

$$\frac{p}{1-qe^t} \qquad (1)$$

Here $M_X(t) = (5 - 4e^t)^{-1} \Rightarrow 5^{-1}\left[1 - \frac{4}{5}e^t\right]^{-1}$ (2)

Company (1) & (2) we get $q = \frac{4}{5}$; $p = \frac{1}{5}$

$\therefore P(X = x) = pq^x, \quad x = 0, 1, 2, 3, ...$

$$= \left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^x$$

$$P(X = 5) = \left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^5 = \frac{4^5}{5^6}$$

Example:33

Suppose that a trainee soldier shoots a target in an independent fashion. If the probability that the target is shot on any one shot is 0.7.

(i) What is the probability that the target would be hit in 10th attempt?
(ii) What is the probability that it takes him less than 4 shots?
(iii) What is the probability that it takes him an even number of shots?
(iv) What is the average no. of shots needed to hit the target?

Solution

Let X denote the no. of shots needed to hit the target and X follows geometric distribution with p.m.f $p[X=x]=pq^{x-1}$, $x=1,2,....$

Given p=0.7 , and q = 1- p = 0.3

(i) $p[x=10]=(0.7)(0.3)^{10-1}=0.0000138$

(ii)
$$p[x<4]=p(x=1)+p(x=2)+p(x=3)$$
$$=(0.7)(0.3)^{1-1}+(0.7)(0.3)^{2-1}+(0.7)(0.3)^{3-1}$$
$$=0.973$$

(iii) $p[x\ is\ an\ even\ number]=p(x=2)+p(x=4)+.....$

$$=(0.7)(0.3)^{2-1}+(0.7)(0.3)^{4-1}+.....$$
$$=(0.7)(0.3)\left[1+(0.3)^2+(0.3)^4.....\right]$$
$$=0.21\left[1+\left((0.3)^2\right)+\left((0.3)^2\right)^2+.....\right]$$
$$=0.21\left[1-(0.3)^2\right]^{-1}=(0.21)(0.91)^{-1}$$
$$=\frac{0.21}{0.91}=0.231$$

(iv) Average no. of shots = $E(X)=\frac{1}{p}=\frac{1}{0.7}=1.4286$

2.10 CONTINUOUS DISTRIBUTIONS

If 'X' is a continuous random variable then we have the following distribution

1. Uniform (Rectangular Distribution)
2. Exponential Distribution
3. Normal Distribution

2.10.1 Uniform Distribution (Rectangular Distribution)

Defintion: A random variable X is set to follow uniform distribution if it has the pdf

$$f(x)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & \text{otherwise}\end{cases}$$

*** To find MGF**

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$= \int_a^b e^{tx} \frac{1}{b-a} dx$$

$$= \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_b^a$$

$$= \frac{1}{(b-a)t}\left[e^{bx} - e^{at}\right]$$

$\therefore$ The MGF of uniform distribution is

$$M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$$

* To find Mean and Variance

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

$$= \int_a^b b_x \frac{1}{b-a} dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{\left(\frac{x^2}{2}\right)_a^b}{b-a}$$

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

Mean $\mu_1' = \frac{a+b}{2}$

Putting $r = 2$ in (A), we get

$$\mu_2' = \int_a^b x^2 f(x)dx = \int_a^b \frac{x^2}{b-a} dx$$

$$= \frac{a^2+ab+b^2}{3}$$

$\therefore$ Variance $= \mu_2' - \mu_1'^2$

$$= \frac{b^2+ab+b^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}$$

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

Example:34

If X is uniformly distributed over (-α,α), find α so that

(i) P(X>1) = 1/3

(ii) P(|X| < 1) = P(|X| > 1)

Solution

If X is uniformly distributed in (-α, α), then its p.d.f. is

$$f(x) = \begin{cases} \frac{1}{2\alpha} & -\alpha < x < \alpha \\ 0 & \text{otherwise} \end{cases}$$

(i) $P(X>1) = 1/3$

$$\int_1^{\alpha} f(x)dx = 1/3$$

$$\int_1^{\alpha} \frac{1}{2\alpha}dx = 1/3$$

$$\frac{1}{2\alpha}(x)_1^{\alpha} = 1/3 \qquad \Rightarrow \frac{1}{2\alpha}(\alpha - 1) = 1/3$$

$$\alpha = 3$$

(ii) $P(|X| < 1) = P(|X| > 1) = 1 - P(|X| < 1)$

$$P(|X| < 1) + P(|X| < 1) = 1$$

$$2\,P(|X| < 1) = 1$$

$$2\,P(-1 < X < 1) = 1$$

$$2\int_{-1}^{1} f(x)dx = 1$$

$$2\int_{-1}^{1} \frac{1}{2\alpha}\,dx = 1$$

$$\Rightarrow \alpha = 2$$

Example:35

Show that for the uniform distribution $f(x) = \frac{1}{2a}$, $-a < x < a$, the mgf about origin is $\frac{\sinh at}{at}$.

Solution: Given $f(x) = \frac{1}{2a}$, $-a < x < a$

MGF

$$M_x(t) = E\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-a}^{a} e^{tx} \frac{1}{2a} dx = \frac{1}{2a}\int_{-a}^{a} e^{tx} dx = \frac{1}{2a}\left[\frac{e^{tx}}{t}\right]_{-a}^{a}$$

$$= \frac{1}{2at}\left[e^{at} - e^{-at}\right] = \frac{1}{2at} 2\sinh at$$

$$= \frac{\sinh at}{at}$$

Note:

1. The distribution function F(x) is given by

$$F(x) = \begin{cases} 0 & -\alpha < x < \alpha \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & b < x < \infty \end{cases}$$

2. The p.d.f. of a uniform variate 'X' in (-a, a) is given by

$$F(x) = \begin{cases} \frac{1}{2a} & -a < x < a \\ 0 & \text{otherwise} \end{cases}$$

2.10.2 THE EXPONENTIAL DISTRIBUTION

Definition: A continuous random variable 'X' is said to follow an exponential distribution with parameter $\lambda>0$ if

The density function of exponential distribution is given by

$$f(x)=\lambda e^{-\lambda x}, x \geq 0$$

$$Mean = E[x] = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$= \int_{0}^{\infty} x\lambda e^{-\lambda x}\,dx = \lambda \int_{0}^{\infty} x e^{-\lambda x}\,dx$$

$$= \lambda\left[\frac{-xe^{-\lambda x}}{\lambda} - \frac{e^{-\lambda x}}{\lambda^2}\right]_0^{\infty}$$

$$= \lambda\left[(0-0)-\left(0-\frac{1}{\lambda^2}\right)\right] = \lambda\left(\frac{1}{\lambda^2}\right) = \frac{1}{\lambda}$$

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

$$E[x^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

$$= \int_{0}^{\infty} x^2\lambda e^{-\lambda x}\,dx = \lambda \int_{0}^{\infty} x^2 e^{-\lambda x}\,dx$$

$$= \lambda\left[\frac{-x^2e^{-\lambda x}}{\lambda} - \frac{2xe^{-\lambda x}}{\lambda^2} - \frac{2e^{-\lambda x}}{\lambda^3}\right]_0^{\infty}$$

$$= \lambda\left[(0-0-0)-\left(0-0-\frac{2}{\lambda^3}\right)\right] = \lambda\left(\frac{2}{\lambda^3}\right) = \frac{2}{\lambda^2}$$

$$Variance = E(x^2) - [E(x)]^2$$

$$= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

PROBLEM ON EXPONENTIAL DISTRIBUTION

Example:36

Let 'X' be a random variable with p.d.f

$$F(x) = \begin{cases} \frac{1}{3}e^{\frac{-x}{3}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find 1) $P(X > 3)$ 2) MGF of 'X'

Solution

WKT the exponential distribution is

$F(x) = \lambda e^{-\lambda x}, \quad x > 0$

Here $\lambda = \frac{1}{3}$

$P(x>3) = \int_3^\infty f(x)dx = \int_3^\infty \frac{1}{3}e^{-\frac{x}{3}}dx$

$P(X>3) = e^{-1}$

MGF is $M_X(t) = \frac{\lambda}{\lambda - t}$

$$= \frac{\frac{1}{3}}{\frac{1}{3} - t} = \frac{\frac{1}{3}}{\frac{1-3t}{3}} = \frac{1}{1-3t}$$

$M_X(t) = \frac{1}{1-3t}$

Example:37

The time (in hours) required to repair a machine is exponentially distributed with parameter $\lambda = \frac{1}{2}$

(a)What is the probability that the repair time exceeds 2 hours?

(b)What is the conditional probability that a repair takes atleast 11 hours given that Its direction exceeds 8 hours?

Solution:

If X represents the time to repair the machine, the density function of X is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$= \frac{1}{2}e^{-x/2} \quad x \geq 0$$

$$(a)\quad P(x>2)=\int_2^{\infty} f(x)\,dx=\int_2^{\infty}\lambda e^{-\lambda x}\,dx$$

$$=\int_2^{\infty}\frac{1}{2}e^{\frac{-x}{2}}\,dx=\frac{1}{2}\left[\frac{e^{-\frac{x}{2}}}{-\frac{1}{2}}\right]_2^{\infty}$$

$$=-\left[0-e^{-1}\right]=0.3679$$

$$(b)\quad P[x\geq 11/x>8]=P[x>3]$$

$$=\int_3^{\infty} f(x)\,dx=\int_3^{\infty}\lambda e^{-\lambda x}\,dx$$

$$=\int_3^{\infty}\frac{1}{2}e^{\frac{-x}{2}}\,dx=\frac{1}{2}\left[\frac{e^{-\frac{x}{2}}}{-\frac{1}{2}}\right]_3^{\infty}$$

$$=-\left[0-e^{-\frac{3}{2}}\right]=e^{-\frac{3}{2}}=0.2231$$

Establish the memory less property of exponential distribution.

If X is exponentially distributed, then
$P(X > s+t / x > s) = P(X > t)$, for any s, t > 0.

Solution:

If X is exponentially distributed, then $P[x>s+t/x>s]=P[x>t]$ *for any* $s,t>0$

The p.d.f of exponential distribution is given by

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

$$P(x>k)=\int_k^{\infty} f(x)\,dx$$

$$= \int_k^{\infty} \lambda e^{-\lambda x}\, dx = \lambda \left[\frac{e^{-\lambda x}}{-\lambda} \right]_k^{\infty}$$

$$= -\left[0 - e^{-\lambda k}\right] = e^{-\lambda x} \text{ ---------(1)}$$

$$P\left[x > s+t / x > s\right] = \frac{P\left[x > s+t \text{ and } x > s\right]}{P\left[x > s\right]}$$

$$= \frac{P\left[x > s+t\right]}{P\left[x > s\right]} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P\left[x > t\right]$$

2.10.3 Normal Or Gaussian Distributions

A r.v. X is said to follow normal distribution with mean μ and variance σ^2, denoted by $N(\mu, \sigma)$ if its density function is given by the probability law

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

MOMENT GENERATING FUNCTION:

M.G.F = $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

MEAN AND VARIANCE:

Mean = μ and Variance = σ^2

Properties of Normal distribution are

1. The normal curve is symmetric about a vertical axis through the mean μ.
2. The total area of the curve is equal to 1.
3. The mode which is the value of the random variable for which $f(x)$ is maximum occurs at $x = \mu$.
4. The normal curve approaches asymptotically the horizontal axis as x increases in either direction away from the mean.
5. The curve has its point of inflection at $x = \mu \pm \sigma$ and is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.
6. The moment generating function is $e^{t\mu + \frac{\sigma^2 t^2}{2}}$.

Find the median of Normal distribution.

Solution:

Median is the point which divides the entire distribution into two equal parts.

$$\int_{-\infty}^{M} f(x)dx = \int_{M}^{\infty} f(x)dx = \frac{1}{2}$$

$$\int_{M}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$$

$$put \quad \frac{(x-\mu)^2}{2\sigma^2} = u$$

$$x - \mu = \sqrt{2}\,\sigma\sqrt{u}$$

$$dx = \frac{\sigma}{\sqrt{2}\sqrt{u}} du$$

$$when \quad x = M,\ u = \frac{(M-\mu)^2}{2\sigma^2} \quad and \quad when\ x = \infty,\ u = \infty$$

$$\int_{\frac{(M-\mu)^2}{2\sigma^2}}^{\infty} \frac{1}{\sigma\sqrt{2}\sqrt{\pi}} e^{-u} \frac{\sigma}{\sqrt{2}\sqrt{u}} du = \frac{1}{2}$$

$$\Rightarrow \frac{1}{2\sqrt{\pi}} \int_{\frac{(M-\mu)^2}{2\sigma^2}}^{\infty} e^{-u}\, u^{-\frac{1}{2}} du = \frac{1}{2}$$

$$\Rightarrow \int_{\frac{(M-\mu)^2}{2\sigma^2}}^{\infty} e^{-u}\, u^{-\frac{1}{2}}\, du = \sqrt{\pi}$$

$$\int_{\frac{(M-\mu)^2}{2\sigma^2}}^{\infty} e^{-u}\, u^{-\frac{1}{2}}\, du = \sqrt{\pi} = \Gamma_{\frac{1}{2}} = \int_0^{\infty} u^{\frac{1}{2}-1} e^{-u}\, du$$

Comparing the integrals, we have $\frac{(M-\mu)^2}{2\sigma^2} = 0 \Rightarrow M - \mu = 0$

$$\Rightarrow M = \mu$$

Therefore, Median = μ .

Example:38

An electrical firm manufactures light bulbs that have a life, before burn out, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find (1) the probability that a bulb burns more than 834 hours (2) the probability that bulb burns between 778 and 834 hours.

Solution:

Given $\mu = 800$, $\sigma = 40$

(1) $P[X > 834] = P\left[\frac{X-\mu}{\sigma} > \frac{834-800}{40}\right]$

$= P[z > 0.85]$

$= 0.5 - P[0 < z < 0.85]$

$= 0.5 - 0.3023 = 0.1977$

(2) $P[778 < X < 834] = P\left[\frac{778-800}{40} < \frac{X-\mu}{\sigma} < \frac{834-800}{40}\right]$

$= P[-0.55 < z < 0.85]$

$= P[-0.55 < z < 0] + P[0 < z < 0.85]$

$= P[0 < z < 0.55] + P[0 < z < 0.85]$

$= 0.2088 + 0.3023$

$= 0.5111$

Example:39

The marks obtained by a number of students in a certain subject are approximately normally distributed with mean 65 and standard deviation 5. If 3 students are selected at random from this group, what is the probability that at least one of them have scored above 75?

Solution:

If X represents the marks obtained by the students, X follows the distribution

$N(65, 5)$.

$P[a\ student\ scores\ above 75] = P[X > 75]$

$= P[75 < X < \infty]$

$= P\left[\frac{75-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \infty\right]$

$= P\left[\frac{75-65}{5} < z < \infty\right]$

$= P[2 < z < \infty]$

$= 0.5 - P[0 < z < 2].$

$= 0.5 \quad 0.4772$

$= 0.0228$.

Let $p = P[a\ student\ score\ above\ 75] = 0.0228$

Then $q = 1 - p = 1 - 0.0228 = 0.9772$ *and* $n = 3$

Since p is the same for all the students the number Y of students scoring above 75 follows Binomial distribution.

$$P[atleast\ 1\ student\ score\ above 75] = P[Y \geq 1]$$

$$= 1 - P[Y < 1]$$

$$= 1 - P[Y = 0]$$

$$= 1 - n_{C_0}\, p^0\, q^{n-0}$$

$$= 1 - 3_{C_0}\, q^3$$

Summary

- Basic knowledge about Probability
- Total probability and Baye's theorem
- A random variable is a function from the state space to the set of real numbers.
- The cumulative distribution function (cdf) of X
- Important discrete distributions include the binomial, Poisson and Geometric
- Important continuous distribution includes Uniform, Exponential and Normal distribution.
- The Probability axioms as well as rules and the distribution of discrete and continuous ideas in solving real world problems.

Keywords

Baye's Theorem:

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

PROBABILITY DENSITY FUNCTION:

$$\lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = f(x)$$

EXPECTATION :

$\overline{X} = E(X) = \sum x_i p_i$ (Discrete random variable)

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

MOMENTS ABOUT ORIGIN (OR RAW MOMENT):

r^{th} moment about origin i.e., $\mu_r' = E(X^r)$

i.e., $\mu_r' = \sum X^r P(X)$, if X is discrete

i.e., $\mu_r' = \int_{-\infty}^{\infty} X^r f(x)dx$, if X is continuous

MOMENT GENERATING FUNCTION (MGF):

$$M_X(t) = E\left(e^{tX}\right) = \sum_x e^{tX} P(x) \text{ , X is discrete.}$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \text{ , X is continuous.}$$

BINOMIAL DISTRIBUTION:

If X is a r.v. such that the p.m.f is $P(X = x) = {}_{n}C_{x}\, p^x\, q^{n-x}$, $x = 0,1,2,\dots,n$ where $p + q = 1$ then X is said to follow the **binomial distributions** with parameters n and p represented as $B(n, p)$.

POISSON DISTRIBUTION:

If X is a r.v. than can assume the values 0,1,2…. ∋ its p.m.f is $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0,1,2,\dots,\lambda > 0$ then X is said to follow a **Poisson distributions** with parameter λ (or) X is said to follow $P(\lambda)$.

GEOMETRIC DISTRIBUTION:

A discrete r.v X, which assumes non – negative values such that its p.m.f is given by $P(X = x) = q^x p$, $x = 0,1,2,\dots,\dots$ where $p + q = 1$ is said to follow Geometric Distribution.

UNIFORM OR RECTANGULAR DISTRIBUTIONS:

A r.v X is said to follow uniform or rectangular distributions over in interval *(a, b)*, if its p.d.f is given by $(x) = \begin{cases} \frac{1}{b-a}, a < x < b \\ 0, \ otherwise \end{cases}$. Here *a* and b *(b > a)* are the parameters of the distribution.

EXPONENTIAL DISTRIBUTION:

A continuous r.v. X is said to follow exponential distributions with parameters $\lambda > 0$, if its p.d.f is given by $f(x) = \begin{cases} \lambda\, e^{-\lambda x}, \ x > 0 \\ 0, \ otherwise \end{cases}$.

NORMAL OR GAUSSIAN DISTRIBUTIONS:

A r.v. X is said to follow normal distribution with mean μ and variance σ^2, denoted by $N(\mu, \sigma)$ if its density function is given by the probability law

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

SELF-ASSESSMENT QUESTIONS

1. It is known that the probability of an item produced by a certain machine will be defective is 0.05. If the produced items are sent to the market in packets of 20, fine the no. of packets containing at least, exactly and atmost 2 defective items in a consignment of 1000 packets using (i) Binomial distribution (ii) Poisson approximation to binomial distribution.
2. The daily consumption of milk in excess of 20,000 gallons is approximately exponentially distributed with 3000 The city has a daily stock of 35,000 gallons. What is the probability that of two days selected at random, the stock is insufficient for both days?
3. The density function of a random variable X is given by f(x)= KX(2-X), $0 \leq X \leq 2$.Find K, mean, variance and rth moment.
4. A binomial variable X satisfies the relation 9P(X=4)=P(X=2) when n=6. Find the parameter p of the Binomial distribution.
5. If X and Y are independent Poisson variates such that P(X=1)=P(X=2) and P(Y=2)=P(Y=3). Find V(X-2Y).

   A discrete random variable has the following probability distribution

X:	0	1	2	3	4	5	6	7	8
P(X)	a	3a	5a	7a	9a	11a	13a	15a	17a

   Find the value of a, P(X<3) and c.d.f of X.

6. In a component manufacturing industry, there is a small probability of 1/500 for any component to be defective. The components are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing (1). No defective. (2). Two defective components in a consignment of 10,000 packets.

7. Suppose that a trainee soldier shoots a target in an independent fashion. If the probability that the target is shot on any one shot is 0.7.
   (i) What is the probability that the target would be hit in 10th attempt?
   (ii) What is the probability that it takes him less than 4 shots?
   (iii) What is the probability that it takes him an even number of shots?
   (iv) What is the average no. of shots needed to hit the target?

8. Starting at 5.00 am every half an hour there is a flight from San Fransisco airport to Los angles. Suppose that none of three planes is completely sold out and that they always have room for passengers. A person who wants to fly to Los angles arrives at a random time between 8.45 am and 9.45 am. Find the probability that she waits (a) Atmost 10 min (b) atleast 15 min.

9. The time (in hours) required to repair a machine is exponentially distributed with parameter $\lambda = \frac{1}{2}$
   (a)What is the probability that the repair time exceeds 2 hours?
   (b)What is the conditional probability that a repair takes atleast 11 hours given that Its direction exceeds 8 hours?

Further Readings

1. Devore, J.L, Probability and Statistics for Engineering and Sciences, Cengage Learning, 8th Edition, New Delhi, 2014.
2. Miller and M. Miller, Mathematical Statistics, Pearson Education Inc., Asia 7th Edition, New Delhi, 2011.
3. Richard Johnson, Miller and Freund's Probability and Statistics for Engineer, Prentice Hall of India Private Ltd., 8th Edition, New Delhi, 2011.

Unit III

TWO DIMENSIONAL RANDOM VARIABLES

CONTENTS

Learning Objectives

In this chapter a student has

- To know the fundamental concepts of joint, marginal and conditional distributions
- To understand covariance and correlation.
- An ability to design a model or a process to meet desired needs within realistic constraints such as environmental conditions

Learning Outcomes

Upon completion of this Unit, students are able to demonstrate a good understanding of:

- To find joint probabilities function in both discrete and Continuous cases.
- Finding Marginal Probability Functions and Conditional Probabilities.
- Finding regression lines and Angle between them
- Finding Covariance and Coefficient of Correlation

Overview

In this Unit. We will see probability function of two variables only. In otherwords, we will deal with only two-dimensional random variables because many situations of interest in engineering can be handled by the theory of two random variables.

3.1 Introduction

In real life, we are often interested in several random variables that are related to each other. For example, suppose that we choose a random family, and we would like to study the number of people in the family, the household income, the ages of the family members, etc. Each of these is a random variable, and we suspect that they are dependent. In this chapter, we develop tools to study joint distributions of random variables. The concepts are similar to what we have seen so far. The only difference is that instead of one random variable, we consider two or more. In this chapter, we will focus on two random variables, but once you understand the theory for two random variables, the extension to nn random variables is straightforward. We will first discuss joint distributions of discrete random variables and then extend the results to continuous random variables.

Definition:

Let S be the sample space. Let X = X(S) & Y = Y(S) be two functions each assigning a real number to each outcome s ∈ S. hen (X, Y) is a two-dimensional random variable.

Joint Distributions: (Two Random Variables)

Remember that for a discrete random variable X, we define the PMF as PX(x)=P(X=x). Now, if we have two random variables X and Y, and we would like to study them jointly, we define the joint probability mass function as follows:

3.2 Two dimensional Discrete Random Variables:

3.2.1 Joint Probability Mass Function:

If For A Two Dimensional Random Variable (X, Y) The Probability That $X=X_i$ & $Y = Y_i$ Is given by $P_{ij}=P(X =x_i, Y=y_j)$ then P_{ij} is called as the joint probability mass function or simply joint probability function of (X,Y) provided it satisfies the following

(i) $P_{ij} \geq 0 \ \forall\, i \,\&\, j$

(ii) $\sum_i \sum_j P(x_i, y_i) = 1$

The function $P(X =x_i, Y=y_i)= P(x_i, y_i)$ is called joint probability function for discrete random variable X and Y is denoted by P_{ij}

3.2.2 Marginal Probability Function [M.P.F]:

The joint PMF contains all the information regarding the distributions of X and Y. This means that, for example, we can obtain PMF of X from its joint PMF with Y.

The marginal probability mass function of X and Y is given by

$$P_{i*} = \sum_i P_{ij} \quad ; \quad P_{*j} = \sum_j P_{ij} \ \text{(discrete)}$$

3.2.3 Joint Cumulative Distributive Function (JCDF)

Remember that, for a random variable X, we define the CDF as $F_X(x)=P(X\leq x)$. Now, if we have two random variables X and Y and we would like to study them jointly, we can define the joint cumulative function as follows:

The joint cumulative distribution function of two random variables X and Y is defined as $F_{XY}(x,y)=P(X\leq x, Y\leq y)$.

$F_{XY}(x,y)=P(X\leq x, Y\leq y)=P(X\leq x)$ and $(Y\leq y))=P((X\leq x)\cap(Y\leq y))$.

3.2.3 Conditional Probability Function:

The conditional probability mass function of X and Y is given by

$$P(X = x_i \ / \ Y = y_j) = \frac{P(X=x_i \cap Y=y_j)}{P(Y=y_j)} = \frac{P_{ij}}{P_{*j}} \ \text{(discrete)}$$

$$P(Y = y_j \ / \ X = x_i) = \frac{P(Y=y_j \cap X=x_i)}{P(X=x_i)} = \frac{P_{ij}}{P_{i*}}$$

Properties Of Joint Distribution:

$(i)\ F[-\infty, y]=0=F[x, -\infty] \text{ and } F[-\infty, \infty]=1$

$(ii)\ P[a<X<b, Y\leq y]=F(b, y)-F(a, y)$

$(iii)\ P[X\leq x, c<Y<d]=F[x, d]-F[x, c]$

$(iv)\ P[a<X<b, c<Y<d]=F[b, d]-F[a, d]-F[b, c]+F[a, c]$

Example: 1

Three balls are drawn at random without replacement from a box containing 2 white, 3 red and 4 black balls. If X denotes the number of white balls drawn and Y denotes the number of red balls drawn, find the joint probability distribution of (X, Y).

Solution:

As there are only 2 white balls in the box, X can take the values 0, 1 and 2

and Y can take the values 0, 1, 2 and 3 since there are only 3 red balls.

$P[X=0\ ,\ Y=0] = P[$drawing 3*balls none of which is white or red*$]$

=P[all three balls drawn are black]

$$= \frac{4c_3}{9c_3} = \frac{1}{21}$$

$$P[X=0\ ,\ Y=1] = P[drawing\ 3\ balls\ 1\ red\ and\ 2\ black\ ] = \frac{3c_1 \times 4c_2}{9c_3} = \frac{3}{14}$$

$$P[X=0\ ,\ Y=2] = P[drawing\ 3\ balls\ 2\ red\ and\ 1\ black\ ] = \frac{3c_2 \times 4c_1}{9c_3} = \frac{1}{7}$$

$$P[X=0\ ,\ Y=3] = P[drawing\ 3\ red\ balls] = \frac{3c_3}{9c_3} = \frac{1}{84}$$

$$P[X=1\ ,\ Y=0] = P[drawing\ 1\ white\ 2\ black\ ] = \frac{2c_1 \times 4c_2}{9c_3} = \frac{1}{7}$$

$$P[X=1\ ,\ Y=1] = P[drawing\ 1\ white\ 1\ red\ and\ 1\ black\ ] = \frac{2c_1 \times 3c_1 \times 4c_1}{9c_3} = \frac{2}{7}$$

$$P[X=1\ ,\ Y=2] = P[drawing\ 1\ white\ 2\ red\ ] = \frac{2c_1 \times 3c_2}{9c_3} = \frac{1}{14}$$

$P[X=1\ ,\ Y=3] = 0$ [since only 3 balls are drawn]

$$P[X=2\ ,\ Y=0] = P[drawing\ 2\ white\ 1\ black\ ] = \frac{2c_2 \times 4c_1}{9c_3} = \frac{1}{21}$$

$$P[X=2\ ,\ Y=1] = P[drawing\ 2\ white\ \ and\ 1\ red\ balls\ ] = \frac{2c_2 \times 3c_1}{9c_3} = \frac{1}{28}$$

$P[X=2\ ,\ Y=2] = 0$ [since only 3 balls are drawn]

$P[X=2\ ,\ Y=3] = 0$ [since only 3 balls are drawn]

The joint probability distribution of $(X\ ,Y)$ may be represented in the form of a table as given below

Y / X	0	1	2	3
0	$\frac{1}{21}$	$\frac{3}{14}$	$\frac{1}{7}$	$\frac{1}{84}$
1	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{1}{14}$	0
2	$\frac{1}{21}$	$\frac{1}{28}$	0	0

Example:2

From the following joint distribution of X and Y find the marginal distributions.

X / Y	0	1	2
0	3/28	9/28	3/28
1	3/14	3/14	0
2	1/28	0	0

Solution

X / Y	0	2	$P_Y(y) = p(Y=y)$
0	3/28 P(0,0)	3/28 P(2,0)	15/28 = $P_y(0)$
1	3/14 P(0, 1)	3/14 P(1,1)	6/14 = $P_y(1)$
2	1/28 P(0,2)	0 P(2,2)	1/28 = $P_y(2)$
$P_X(X) = P(X=x)$	10/28 = 5/14 $P_X(0)$	3/28 $P_X(2)$	1

The marginal distribution of X

$P_X(0) = P(X=0) = p(0,0) + p(0,1) + p(0,2) = 5/14$

$P_X(1) = P(X=1) = p(1,0) + p(1,1) + p(1,2) = 15/28$

$P_X(2) = P(X=2) = p(2,0) + p(2,1) + p(2,2) = 3/28$

Marginal probability function of X is

$$P_X(x) = \begin{cases} \frac{5}{14}, & x = 0 \\ \frac{15}{28}, & x = 1 \\ \frac{3}{28}, & x = 2 \end{cases}$$

The marginal distribution of Y

$P_Y(0) = P(Y=0) = p(0,0) + p(1,0) + p(2,0) = 15/28$

$P_Y(1) = P(Y=1) = p(0,1) + p(2,1) + p(1,1) = 3/7$

$P_Y(2) = P(Y=2) = p(0,2) + p(1,2) + p(2,2) = 1/28$

Marginal probability function of Y is

$$P_Y(y) = \begin{cases} \frac{15}{28}, & y = 0 \\ \frac{3}{7}, & y = 1 \\ \frac{1}{28}, & y = 2 \end{cases}$$

Example:3

The joint probability mass function of (X, Y) is given by $p(x, y) = k(2x+3y)$, x=0,1,2 , y =1,2,3. Find all the marginal and conditional probability distributions. Also find the probability distribution of $X+Y$.

Solution:

The joint probability distribution of (X, Y) is given below

X \ Y	1	2	3
0	3k	6k	9k
1	5k	8k	11k
2	7k	10k	13k

Since $p(x, y)$ is a probability mass function, we have

$$\sum\sum p(x, y) = 1$$

$$3k+6k+9k+5k+8k+11k+7k+10k+13k = 1$$

$$72k = 1$$

$$k = \frac{1}{72}$$

Marginal probability distribution of X

$$P[X=0] = 3k+6k+9k = 18k = \frac{18}{72} = \frac{1}{4}$$

$$P[X=1] = 5k+8k+11k = 24k = \frac{24}{72} = \frac{1}{3}$$

$$P[X=0] = 7k+10k+13k = 30k = \frac{30}{72} = \frac{5}{12}$$

Marginal probability distribution of Y

$$P[Y=1] = 3k+5k+7k = 15k = \frac{15}{72} = \frac{5}{24}$$

$$P[Y=2] = 6k+8k+10k = 24k = \frac{24}{72} = \frac{1}{3}$$

$$P[Y=3] = 9k+11k+13k = 33k = \frac{33}{72} = \frac{11}{24}$$

Conditional distribution of X given $_{y=1}$

$$P[X=0/Y=1] = \frac{P[X=0\ ,\ Y=1]}{P[Y=1]} = \frac{3k}{15k} = \frac{3}{15} = \frac{1}{5}$$

$$P[X=1/Y=1] = \frac{P[X=1\ ,\ Y=1]}{P[Y=1]} = \frac{5k}{15k} = \frac{5}{15} = \frac{1}{3}$$

$$P[X=2/Y=1] = \frac{P[X=2\ ,\ Y=1]}{P[Y=1]} = \frac{7k}{15k} = \frac{7}{15}$$

Conditional distribution of X given $Y=2$

$$P[X=0/Y=2] = \frac{P[X=0\ ,\ Y=2]}{P[Y=2]} = \frac{6k}{24k} = \frac{6}{24} = \frac{1}{4}$$

$$P[X=1/Y=2] = \frac{P[X=1\ ,\ Y=2]}{P[Y=2]} = \frac{8k}{24k} = \frac{8}{24} = \frac{1}{3}$$

$$P[X=2/Y=2] = \frac{P[X=2\ ,\ Y=2]}{P[Y=2]} = \frac{10k}{24k} = \frac{5}{12}$$

Conditional distribution of X given $Y=3$

$$P[X=0/Y=3] = \frac{P[X=0\ ,\ Y=3]}{P[Y=3]} = \frac{9k}{33k} = \frac{9}{33} = \frac{3}{11}$$

$$P[X=1/Y=3] = \frac{P[X=1\ ,\ Y=3]}{P[Y=3]} = \frac{11k}{33k} = \frac{11}{33} = \frac{1}{3}$$

$$P[X=2/Y=3] = \frac{P[X=2\ ,\ Y=3]}{P[Y=3]} = \frac{13k}{33k} = \frac{13}{33}$$

Conditional distribution of Y given $X=0$

$$P[Y=1/X=0] = \frac{P[X=0\ ,\ Y=1]}{P[X=0]} = \frac{3k}{18k} = \frac{3}{18} = \frac{1}{6}$$

$$P[Y=2/X=0] = \frac{P[X=0\ ,\ Y=2]}{P[X=0]} = \frac{6k}{18k} = \frac{6}{18} = \frac{1}{3}$$

$$P[Y=3/X=0]=\frac{P[X=0\ ,\ Y=3]}{P[X=0]}=\frac{9k}{18k}=\frac{9}{18}=\frac{1}{2}$$

Conditional distribution of Y given $X=1$

$$P[Y=1/X=1]=\frac{P[X=1\ ,\ Y=1]}{P[X=1]}=\frac{5k}{24k}=\frac{5}{24}$$

$$P[Y=2/X=1]=\frac{P[X=1\ ,\ Y=2]}{P[X=1]}=\frac{8k}{24k}=\frac{8}{24}=\frac{1}{3}$$

$$P[Y=3/X=1]=\frac{P[X=1\ ,\ Y=3]}{P[X=1]}=\frac{11k}{24k}=\frac{11}{24}$$

Conditional distribution of Y given $X=2$

$$P[Y=1/X=2]=\frac{P[X=2\ ,\ Y=1]}{P[X=2]}=\frac{7k}{30k}=\frac{7}{30}$$

$$P[Y=2/X=2]=\frac{P[X=2\ ,\ Y=2]}{P[X=2]}=\frac{10k}{30k}=\frac{10}{30}=\frac{1}{3}$$

$$P[Y=3/X=2]=\frac{P[X=2\ ,\ Y=3]}{P[X=2]}=\frac{13k}{30k}=\frac{13}{30}$$

Probability distribution of $(X+Y)$

$X+Y$	p
1	$p_{01}=3k=\frac{3}{72}$
2	$p_{02}+p_{11}=6k+5k=11k=\frac{11}{72}$
3	$p_{03}+p_{12}+p_{21}=9k+8k+7k=24k=\frac{24}{72}$
4	$p_{13}+p_{22}=11k+10k=21k=\frac{21}{72}$
5	$p_{23}=13k=\frac{13}{72}$

3.3 Two dimensional Continuous Random variables

3.3.1 Joint Probability Density Function:

Here, we will define jointly continuous random variables. Basically, two random variables are jointly continuous if they have a joint probability density function as defined below.

If (X,Y) is a two dimensional continuous random variable such that

$P\left[x-\frac{dx}{2}\le X\le x+\frac{dx}{2},\ y-\frac{dy}{2}\le Y\le y+\frac{dy}{2}\right]=f(x,y)dx\,dy$, then $f(x,y)$ is called the joint p.d.f of (X,Y), provided $f(x,y)$ satisfies the following conditions

$(i)\ f(x,y)\ge 0\ for\ all\ (x,y)\in R$

$(ii)\iint_R f(x,y)dx\,dy=1$

3.3.2 Marginal probability functions

The m.p.f of X and Y is given by

$$f_x(x)=f(x)=\int_{-\infty}^{\infty}f(x,y)dy \quad \text{(continuous)}$$

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

3.3.3 Conditioning by Another Random Variable:

If X and Y are two jointly continuous random variables, and we obtain some information regarding Y,

The conditional PDF of Y given X=x is given by

$$f_{X/Y}=\frac{f(x,y)}{f_Y(y)}$$

The conditional PDF of Y given X=x is given by

$$f_{Y/X}=\frac{f(x,y)}{f_x(x)}$$

Independent Random Variables:

When two jointly continuous random variables are independent,

$$f(x,y)=f(x)f(y)$$

Example:4

The j.d.f of the random variables X and Y is given

$$f(x,y)=\begin{cases}8xy, & 0<x<1, \quad 0<y<x\\ 0, & \text{otherwise}\end{cases}$$

Find (i) $f_X(x)$ (ii) $f_Y(y)$ (iii) $f(y/x)$

Solution

We know that

(i) The marginal pdf of 'X' is

$$f_X(x) = f(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^x 8xy\,dy = 4x^3$$

$$f(x)=4x^3,\ 0<x<1$$

(ii) The marginal pdf of 'Y' is

$$f_Y(y) = f(y) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 8xy\,dy = 4y$$

$$f(y)=4y,\ 0<y<\alpha$$

(iii) We know that

$$f(y/x) = \frac{f(x,y)}{f(x)}$$

$$=\frac{8xy}{4x^3} = \frac{2y}{x^2},\ 0<y<x,\ 0<x<1$$

Example:5

Given the joint pdf of

$$f(x,y)=\begin{cases}e^{-(x+y)}, & x>0,\ y>0\\ 0 & ,elsewhere\end{cases}$$

Find the marginal densities of *X* and *Y*. Are *X* and *Y* independent?

Solution:

Marginal density of X is

$$f_X(x)=\int f(x,y)\,dy$$

$$=\int_0^{\infty} e^{-x}e^{-y}\,dy = e^{-x}\int_0^{\infty} e^{-y}\,dy = e^{-x}\left(-e^{-y}\right)_0^{\infty}$$

$$=-e^{-x}(0-1)=e^{-x},\ x>0$$

Marginal density of Y is

$$f_Y(y)=\int f(x,y)\,dx$$

$$=\int_0^\infty e^{-x}\,e^{-y}\,dx=e^{-y}\int_0^\infty e^{-x}\,dx=e^{-y}\left(-e^{-x}\right)_0^\infty$$

$$=-e^{-y}(0-1)=e^{-y}\ ,y>0$$

$$f_X(x).f_Y(y)=e^{-x}.e^{-y}=e^{-(x+y)}=f_{XY}(x,y)$$

Therefore X and Y are independent.

Example:6

$$f(x,y)=k\left(x^3y+xy^3\right)\ ,0\le x\le 2;0\le y\le 2$$

Find the value of k and marginal and conditional density functions.

Solution:

Given $f(x,y)$ is the joint pdf, we have

$$\iint f(x,y)dx\,dy=1$$

$$k\int_0^2\int_0^2\left(x^3y+xy^3\right)dx\,dy=1$$

$$k\int_0^2\left[y\left(\frac{x^4}{4}\right)_0^2+y^3\left(\frac{x^2}{2}\right)_0^2\right]dy=1$$

$$k\int_0^2\left(4y+2y^3\right)dy=1$$

$$k\left[4\left(\frac{y^2}{2}\right)_0^2+2\left(\frac{y^4}{4}\right)_0^2\right]=1$$

$$k[8+8]=1$$

$$16k=1$$

$$k=\frac{1}{16}$$

Therefore, $f(x,y)=\frac{1}{16}\left(x^3y+xy^3\right);0\le x\le 2,0\le y\le 2$

Marginal density of X is

$$f_X(x)=\int f(x,y)\,dy=\int_0^2\frac{1}{16}\left(x^3y+xy^3\right)dy$$

$$=\frac{1}{16}\left[x^3\left(\frac{y^2}{2}\right)_0^2+x\left(\frac{y^4}{4}\right)_0^2\right]=\frac{1}{16}\left[\frac{x^3}{2}(4-0)+\frac{x}{4}(16-0)\right]$$

$$=\frac{1}{16}\left[2x^3+4x\right]=\frac{x^3+2x}{8},\ 0\le x\le 2$$

Marginal density of Y is

$$f_Y(y)=\int f(x,y)\,dx=\int_0^2\frac{1}{16}\left(x^3y+xy^3\right)dx$$

$$=\frac{1}{16}\left[y\left(\frac{x^4}{4}\right)_0^2+y^3\left(\frac{x^2}{2}\right)_0^2\right]=\frac{1}{16}\left[\frac{y}{4}(16-0)+\frac{y^3}{2}(4-0)\right]$$

$$=\frac{1}{16}\left[4y+2y^3\right]=\frac{y^3+2y}{8},\ 0\le y\le 2$$

Conditional density of X given Y is

$$f_{X/Y}(x/y)=\frac{f(x,y)}{f_Y(y)}=\frac{\frac{1}{16}\left(x^3y+xy^3\right)}{\frac{y^3+2y}{8}}=\frac{8}{16}.\frac{y\left(x^3+xy^2\right)}{y\left(y^2+2\right)}$$

$$f_{X/Y}(x/y)=\frac{\left(x^3+xy^2\right)}{2\left(y^2+2\right)},\ 0\le x\le 2$$

Conditional density of Y given X is

$$f_{Y/X}(y/x)=\frac{f(x,y)}{f_X(x)}=\frac{\frac{1}{16}\left(x^3y+xy^3\right)}{\frac{x^3+2x}{8}}=\frac{8}{16}.\frac{x\left(x^2y+y^3\right)}{x\left(x^2+2\right)}$$

$$f_{Y/X}(y/x)=\frac{\left(x^2y+y^3\right)}{2\left(x^2+2\right)};0\le y\le 2.$$

Example:7

Given the joint pdf of (X,Y) as

$$f(x,y)=\begin{cases}8xy\ ;\ 0<x<y<1\\ 0\qquad ,otherwise\end{cases}$$

Find the marginal and conditional probability density functions of X and Y. Are X and Y are independent?

Solution:

Marginal density of X is

$$f_X(x)=\int f(x,y)\,dy=\int_x^1 8xy\,dy=8x\int_x^1 y\,dy=8x\left(\frac{y^2}{2}\right)_x^1$$

$$=4x(1-x^2),\ 0<x<1$$

Marginal density of Y is

$$f_Y(y)=\int f(x,y)\,dx=\int_0^y 8xy\,dx=8y\int_0^y x\,dx=8y\left(\frac{x^2}{2}\right)_0^y$$

$$=4y(y^2-0)=4y^3\ ,\ 0<y<1$$

$$f_X(x).f_Y(y)=4x(1-x^2).4y^3\neq 8xy\neq f_{XY}(x,y)$$

Therefore X and Y are not independent.

Conditional density of X given Y is

$$f_{X/Y}(x/y)=\frac{f(x,y)}{f_Y(y)}=\frac{8xy}{4y^3}=\frac{2x}{y^2},\ 0<x<y$$

Conditional density of Y given X is

$$f_{Y/X}(y/x)=\frac{f(x,y)}{f_X(x)}=\frac{8xy}{4x(1-x^2)}=\frac{2y}{1-x^2},\ x<y<1$$

Example:8

Given $f_{XY}(x,y)=\begin{cases}cx(x-y)\ ;0<x<2,-x<y<x\\ 0\qquad\qquad ;otherwise\end{cases}$ **(1) Evaluate c, find (2)** $f_X(x)$ **(3)** $f_{Y/X}(y/x)$ **and (4)** $f_Y(y)$**.**

Solution:

(1) Given $f(x,y)$ is the joint p.d.f, we have

$$\iint f(x,y)\,dx\,dy = 1$$

$$c\int_0^2\int_{-x}^{x}\left(x^2-xy\right)dy\,dx = 1$$

$$c\int_0^2\left[x^2\left(y\right)_{-x}^{x} - x\left(\frac{y^2}{2}\right)_{-x}^{x}\right]dx = 1$$

$$c\int_0^2\left[x^2\left(x-(-x)\right)-\frac{x}{2}\left(x^2-x^2\right)\right]dx = 1$$

$$c\int_0^2\left(2x^3-0\right)dx = 1 \quad \Rightarrow 2c\int_0^2 x^3\,dx = 1 \quad \Rightarrow 2c\left[\frac{x^4}{4}\right]_0^2 = 1$$

$$\frac{c}{2}\left[16-0\right] = 1 \quad \Rightarrow 8c = 1 \quad \Rightarrow c = \frac{1}{8}$$

Therefore , $f(x,y) = \frac{1}{8}\left(x^2-xy\right); 0<x<2, -x<y<x$

(2) $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{8}\int_{-x}^{x}\left(x^2-xy\right)dy = \frac{1}{8}\left[x^2\left(y\right)_{-x}^{x} - x\left(\frac{y^2}{2}\right)_{-x}^{x}\right]$

$$= \frac{1}{8}\left[x^2\left(x-(-x)\right)-\frac{x}{2}\left(x^2-x^2\right)\right] = \frac{1}{8}\left[x^2\left(2x\right)-0\right] = \frac{x^3}{4}, 0<x<2.$$

(3) $f_{Y/X}(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{8}\left(x^2-xy\right)}{\frac{x^3}{4}} = \frac{4}{8}\frac{x(x-y)}{x^3} = \frac{x-y}{2x^2}, -x<y<x$

(4) $f_Y(y) = \int f(x,y)\,dx$

$$= \begin{cases} \frac{1}{8}\int_{-y}^{2}\left(x^2 - xy\right)dx & in \ -2 \le y \le 0 \\ \frac{1}{8}\int_{y}^{2}\left(x^2 - xy\right)dx & in \ 0 \le y \le 2 \end{cases}$$

$$= \begin{cases} \frac{1}{8}\left[\left(\frac{x^3}{3}\right)_{-y}^{2} - y\left(\frac{x^2}{2}\right)_{-y}^{2}\right] & in \quad -2 \le y \le 0 \\ \frac{1}{8}\left[\left(\frac{x^3}{3}\right)_{y}^{2} - y\left(\frac{x^2}{2}\right)_{y}^{2}\right] & in \quad 0 \le y \le 2 \end{cases}$$

$$= \begin{cases} \frac{1}{8}\left[\frac{1}{3}\left(8 + y^3\right) - \frac{y}{2}\left(4 - y^2\right)\right] & in \quad -2 \le y \le 0 \\ \frac{1}{8}\left[\frac{1}{3}\left(8 - y^3\right) - \frac{y}{2}\left(4 - y^2\right)\right] & in \quad 0 \le y \le 2 \end{cases}$$

$$= \begin{cases} \frac{1}{3} - \frac{y}{4} + \frac{5}{48}y^3 & in \quad -2 \le y \le 0 \\ \frac{1}{3} - \frac{y}{4} + \frac{1}{48}y^3 & in \quad 0 \le y \le 2 \end{cases}$$

3.4 Covariance and Correlation

Definition:
Consider two random variables X and Y. Here, we define the covariance between X and Y, written Cov(X,Y). The covariance gives some information about how X and Y are statistically related. Let us provide the definition, then discuss the properties and applications of covariance.

The covariance between X and Y is defined as
Cov(X,Y)=E[XY]−(EX)(EY).
Cov (X, Y) = 0 [If X & Y are independent]

The covariance has the following properties:
1. Cov(X,X)=Var(X)
2. if X and Y are independent then Cov(X,Y)=0
3. Cov(X,Y)=Cov(Y,X)
4. Cov(aX,Y)=aCov(X,Y)
5. Cov(X+c,Y)=Cov(X,Y)

6. Cov(X+Y,Z)=Cov(X,Z)+Cov(Y,Z)
7. More generally,
   $Cov(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i\, b_j Cov(X_i, Y_j)$

CORRELATION

Two variables X and Y are said to be correlated if a change in the value of one of the variables causes a change in the value of the other variable.

Correlation coefficient between two random variables X and Y usually denoted by r(X, Y) or $\rho(X,Y)$ is a numerical measure of linear relationship between them and is defined as

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X . \sigma_Y}, \quad \text{Where Cov (X, Y)} = \frac{1}{n}\sum XY - \overline{X}\,\overline{Y}$$

$$if\ \sigma_x \neq 0, \sigma_y \neq 0$$

This is called Karl Pearson's correlation coefficient.

* Limits of correlation coefficient

$-1 \leq r \leq 1$.

X & Y independent, $\therefore$ r(X, Y) = 0.

Note : Types of correlation based on 'r'.

Values of 'r'	Correlation is said to be
r = 1	perfect and positive
0<r<1	positive
-1<r<0	negative
r = 0	Uncorrelated

Example:9

Let X and Y be any two random variables and a,b be constants. Prove that $Cov(aX, bY) = ab\,cov(X, Y)$.

Solution: $Cov(X,Y) = E[XY] - E[X]E[Y]$

$$Cov(aX, bY) = E[aX\ bY] - E[aX]E[bY]$$

$$= ab\,E[XY] - a\,E[X]\,b\,E[Y] = ab[E(XY) - E(X)E(Y)]$$

$$= ab\,Cov(X,Y).$$

Example:10

If $Y=-2X+3$, find $Cov(X,Y)$.

Solution: $Cov(X,Y)=E[XY]-E[X]E[Y]$

$$=E[X(-2X+3)-E[X]E[-2X+3]]$$

$$=E[-2X^2+3X]-E[X](-2E[X]+3)$$

$$=-2E[X^2]+3E[X]+2(E[X])^2-3E[X]$$

$$=-2[E[X^2]-(E[X])^2]=-2Var\,X.$$

Example:11

If $f(x,y)=\begin{cases}2-x-y\ , & 0\le x\le 1,\ 0\le y\le 1\\ 0 & ;elsewhere\end{cases}$

is the joint pdf of the random variables X and Y, find the correlation co-efficient of X and Y.

Solution:

$$\text{Correlation co-efficient}=\frac{E[XY]-E[X]E[Y]}{\sigma_X\,\sigma_Y}$$

$$E[X]=\iint x\,f(x,y)\,dx\,dy$$

$$=\int_0^1\int_0^1 x(2-x-y)\,dx\,dy=\int_0^1\int_0^1 (2x-x^2-xy)\,dx\,dy$$

$$=\int_0^1\left[2\left(\frac{x^2}{2}\right)_0^1-\left(\frac{x^3}{3}\right)_0^1-y\left(\frac{x^2}{2}\right)_0^1\right]dy=\int_0^1\left[1-\frac{1}{3}-\frac{y}{2}\right]dy=\int_0^1\left(\frac{2}{3}-\frac{y}{2}\right)dy$$

$$=\frac{2}{3}(y)_0^1-\frac{1}{2}\left(\frac{y^2}{2}\right)_0^1=\frac{2}{3}-\frac{1}{4}=\frac{5}{12}$$

$$E[Y]=\iint y\,f(x,y)\,dx\,dy$$

$$=\int_0^1\int_0^1 y(2-x-y)\,dx\,dy=\int_0^1\int_0^1 (2y-xy-y^2)\,dx\,dy$$

$$= \int_0^1 \left[2y\,(x)_0^1 - y\left(\frac{x^2}{2}\right)_0^1 - y^2\,(x)_0^1 \right] dy$$

$$= \int_0^1 \left[2y - \frac{y}{2} - y^2 \right] dy = \int_0^1 \left(\frac{3}{2}y - y^2 \right) dy$$

$$= \frac{3}{2}\left(\frac{y^2}{2}\right)_0^1 - \left(\frac{y^3}{3}\right)_0^1 = \frac{3}{4} - \frac{1}{3} = \frac{5}{12}$$

$$E\left[X^2\right] = \iint x^2 f(x,y)\,dx\,dy$$

$$= \int_0^1 \int_0^1 x^2 (2 - x - y)\,dx\,dy = \int_0^1 \int_0^1 \left(2x^2 - x^3 - yx^2\right) dx\,dy$$

$$= \int_0^1 \left[2\left(\frac{x^3}{3}\right)_0^1 - \left(\frac{x^4}{4}\right)_0^1 - y\left(\frac{x^3}{3}\right)_0^1 \right] dy = \int_0^1 \left[\frac{2}{3} - \frac{1}{4} - \frac{y}{3} \right] dy = \int_0^1 \left(\frac{5}{12} - \frac{y}{3} \right) dy$$

$$= \frac{5}{12}(y)_0^1 - \frac{1}{3}\left(\frac{y^2}{2}\right)_0^1 = \frac{5}{12} - \frac{1}{6} = \frac{1}{4}$$

$$E\left[Y^2\right] = \iint y^2 f(x,y)\,dx\,dy$$

$$= \int_0^1 \int_0^1 y^2 (2 - x - y)\,dx\,dy = \int_0^1 \int_0^1 \left(2y^2 - xy^2 - y^3\right) dx\,dy$$

$$= \int_0^1 \left[2y^2\,(x)_0^1 - y^2\left(\frac{x^2}{2}\right)_0^1 - y^3\,(x)_0^1 \right] dy = \int_0^1 \left[2y^2 - \frac{y^2}{2} - y^3 \right] dy$$

$$= \int_0^1 \left(\frac{3}{2}y^2 - y^3 \right) dy = \frac{3}{2}\left(\frac{y^3}{3}\right)_0^1 - \left(\frac{y^4}{4}\right)_0^1 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

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

$$= \frac{1}{4} - \left(\frac{5}{12}\right)^2 = \frac{1}{4} - \frac{25}{144} = \frac{11}{144}$$

$$\sigma_X = \frac{\sqrt{11}}{12}$$

$$\sigma_Y{}^2 = E\left[Y^2\right] - (E[Y])^2$$

$$= \frac{1}{4} - \left(\frac{5}{12}\right)^2 = \frac{1}{4} - \frac{25}{144} = \frac{11}{144}$$

$$\sigma_Y = \frac{\sqrt{11}}{12}$$

$$E[XY] = \iint xy\, f(x,y)\,dx\,dy$$

$$= \int_0^1 \int_0^1 xy\,(2-x-y)\,dx\,dy = \int_0^1 \int_0^1 \left(2xy - x^2y - xy^2\right)dx\,dy$$

$$= \int_0^1 \left[2y\left(\frac{x^2}{2}\right)_0^1 - y\left(\frac{x^3}{3}\right)_0^1 - y^2\left(\frac{x^2}{2}\right)_0^1 \right] dy = \int_0^1 \left[y - \frac{y}{3} - \frac{y^2}{2} \right] dy$$

$$= \int_0^1 \left(\frac{2}{3}y - \frac{1}{2}y^2\right)dy = \frac{2}{3}\left(\frac{y^2}{2}\right)_0^1 - \frac{1}{2}\left(\frac{y^3}{3}\right)_0^1 = \frac{1}{3} - \frac{1}{6} = \frac{1}{6}$$

$$Corr(X,Y) = \frac{\frac{1}{6} - \left(\frac{5}{12}\right)\left(\frac{5}{12}\right)}{\frac{\sqrt{11}}{12}\cdot\frac{\sqrt{11}}{12}} = \frac{\frac{1}{6} - \frac{25}{144}}{\frac{11}{144}} = \frac{\frac{-1}{144}}{\frac{11}{144}} = -\frac{1}{11}$$

Example:12

If the independent random variables X and Y have the variances 36 and 16 respectively, find the correlation co-efficient between $X+Y$ and $X-Y$.

Solution:

Let $U = X+Y$ and $V = X-Y$

Given $Var\,X = \sigma_X^{\,2} = 36 \Rightarrow \sigma_X = 6$

$Var\,Y = \sigma_Y^{\,2} = 16 \Rightarrow \sigma_Y = 4$

Correlation co-efficient $= \rho_{UV} = \dfrac{E[UV] - E[U]E[V]}{\sigma_U\,\sigma_V}$

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

$$E[V] = E[X-Y] = E[X] - E[Y]$$

$$E[UV] = E[(X+Y)(X-Y)] = E[X^2 - Y^2] = E[X^2] - E[Y^2]$$

$$E[U^2] = E[(X+Y)^2] = E[X^2 + 2XY + Y^2]$$

$$= E[X^2] + 2E[XY] + E[Y^2]$$

$$= E[X^2] + 2E[X]E[Y] + E[Y^2]$$

$$E[V^2] = E[(X-Y)^2] = E[X^2 - 2XY + Y^2]$$

$$= E[X^2] - 2E[XY] + E[Y^2]$$

$$= E[X^2] - 2E[X]E[Y] + E[Y^2]$$

$$E[U]E[V] = (E[X] + E[Y])(E[X] - E[Y]) = (E[X])^2 - (E[Y])^2$$

$$\sigma_U{}^2 = E[U^2] - (E[U])^2$$

$$= (E[X^2] + 2E[X]E[Y] + E[Y^2]) - (E[X] + E[Y])^2$$

$$= E[X^2] + 2E[X]E[Y] + E[Y^2] - (E[X])^2 - 2E[X]E[Y] - E[Y^2]$$

$$= \sigma_X{}^2 + \sigma_Y{}^2 = 36 + 16 = 52$$

$$\sigma_U = \sqrt{52}$$

$$\sigma_V{}^2 = E[V^2] - (E[V])^2$$

$$= (E[X^2] - 2E[X]E[Y] + E[Y^2]) - (E[X] - E[Y])^2$$

$$= E[X^2] - 2E[X]E[Y] + E[Y^2] - (E[X])^2 + 2E[X]E[Y] - E[Y^2]$$

$$= [E[X^2] - (E[X])^2] + [E[Y^2] - (E[Y])^2]$$

$$= \sigma_X{}^2 + \sigma_Y{}^2 = 36 + 16 = 52$$

$$\sigma_V = \sqrt{52}$$

$$\rho_{UV} = \frac{[E[X^2] - E[Y^2]] - [(E[X])^2 - (E[Y])^2]}{\sqrt{52}.\sqrt{52}}$$

$$= \frac{[E[X^2] - (E[X])^2] - [E[Y^2] - (E[Y])^2]}{52}$$

$$= \frac{\sigma_X{}^2 - \sigma_Y{}^2}{52}$$

3.5 REGRESSION

Regression is the study of the relationship between the variable. When there is a linear relationship suggested by the scatter diagram, then this line is called the line of regression.

If X is taken as independent variable and Y is the dependent variable, then the line of best fit to the set of observed values(x_i,y_i), i=1,2,3,…,n of (X,Y) by the method of least square is called the regression line of Y on X.

Lines of Regression: The line of Regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \text{ where } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

The line of Regression of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y}) \text{ where } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$r^2 = b_{YX}b_{XY}$$

$$r = \pm\sqrt{b_{YX}b_{XY}}$$

$if\ b_{YX}\ and\ b_{XY}$ are positive, then r is positive and $r = \sqrt{b_{YX}b_{XY}}$

$if\ b_{YX}\ and\ b_{XY}$ are negative, then r is negative and $r = -\sqrt{b_{YX}b_{XY}}$

* Angle between two lines of Regression.

$$\tan\theta = \frac{1-r^2}{r}\left(\frac{\sigma_y\sigma_x}{\sigma_{x^2}+\sigma_{y^2}}\right)$$

Example:13

Calculated the correlation coefficient for the following heights of fathers X and their sons Y

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution:

X	Y	U = X – 68	V = Y – 68	UV	U^2	V^2
65	67	-3	-1	3	9	1
66	68	-2	0	0	4	0
67	65	-1	-3	3	1	9
67	68	-1	0	0	1	0
68	72	0	4	0	0	16
69	72	1	4	4	1	16
70	69	2	1	2	4	1
72	71	4	3	12	16	9

$\sum U = 0 \quad \sum V = 0 \quad \sum UV = 24\ \sum U^2 = 36\ \sum V^2 = 52$

$$\overline{U} = \frac{\Sigma U}{n} = \frac{0}{8} = 0$$

$$\overline{V} = \frac{\Sigma V}{n} = \frac{8}{8} = 1$$

Cov (X, Y) = Cov (U, V)

$$\Rightarrow \frac{\Sigma UV}{n} - \overline{U}\,\overline{V} = \frac{24}{8} - 0 = 3$$

$$\sigma_U = \sqrt{\frac{\Sigma U^2}{n} - \overline{U}^2} = \sqrt{\frac{36}{8} - 0} = 2.121$$

$$\sigma_V = \sqrt{\frac{\Sigma V^2}{n} - \overline{V}^2} = \sqrt{\frac{52}{8} - 1} = 2.345$$

$$\therefore r(X, Y) = r(U, V) = \frac{Cov(U,V)}{\sigma_U . \sigma_V} = \frac{3}{2.121 \times 2.345}$$

r =0.6031

Example:14

From the following data, find

(i) The two-regression equation

(ii) The coefficient of correlation between the marks in Economic and Statistics.

(iii) The most likely marks in statistics when marks in Economic are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	40	46	49	41	36	32	31	30	33	39

Solution

X	Y	$X-\overline{X}=X-32$	$X-\overline{Y}=Y-38$	$(X-X)^2$	$(Y-\overline{Y})^2$	$(X-\overline{X})^2(Y-\overline{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	4	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	09	49	+21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-48
32	39	0	1	0	1	100
320	380	0	0	140	398	-93

Here $\overline{X} = \frac{\Sigma X}{n} = \frac{320}{10} = 32$ and $\overline{Y} = \frac{\Sigma Y}{n} = \frac{380}{10} = 38$

Coefficient of regression of Y on X is

$$b_{YX} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2} = \frac{-93}{140} = -0.6643$$

Coefficient of regression of X on Y is

$$b_{XY} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(Y-\overline{Y})^2} = \frac{-93}{398} = -0.2337$$

Equation of the line of regression of X and Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$
$$X - 32 = -0.2337\,(y - 38)$$
$$X = -0.2337\,y + 0.2337 \times 38 + 32$$
$$X = -0.2337\,y + 40.8806$$

Equation of the line of regression of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$
$$Y - 38 = -0.6643\,(x - 32)$$
$$Y = -0.6643\,x + 38 + 0.6643 \times 32$$
$$= -0.6642\,x + 59.2576$$

Coefficient of Correlation

$$r^2 = b_{YX} \times b_{XY}$$
$$= -0.6643 \times (-0.2337)$$
$$r = 0.1552$$
$$r = \pm\sqrt{0.1552}$$
$$r = \pm\sqrt{0.394}$$

Now we have to find the most likely mark, in statistics (Y) when marks in economics (X) are 30.

$$y = -0.6643\,x + 59.2575$$

Put x = 30, we get

$$y = -0.6643 \times 30 + 59.2536$$
$$= 39.3286$$
$$y \simeq 39$$

Example :15

Let X be a random variable with p.d.f. $f(x) = \frac{1}{2}, -1 \le x \le 1$ and let $Y = x^2$, find the correlation coefficient between X and Y.

Solution:

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

$$= \int_{-1}^{1} x.\frac{1}{2}\,dx \quad = \frac{1}{2}\left(\frac{x^2}{2}\right)_{-1}^{1} \quad = \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\right) = 0$$

$$E(X) = 0$$

$$E(Y) = \int_{-\infty}^{\infty} x^2.f(x)\,dx$$

$$= \int_{-1}^{1} x^2.\frac{1}{2}\,dx \quad = \frac{1}{2}\left(\frac{x^3}{3}\right)_{-1}^{1} \quad = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{3}\right) = \frac{1}{2}.\frac{2}{3} = \frac{1}{3}$$

$$E(XY) = E(XX^2)$$

$$= E(X^3) \qquad = \int_{-\infty}^{\infty} x^3.f(x)\,dx = \left(\frac{x^4}{4}\right)_{-1}^{1} = 0$$

$$E(XY) = 0$$

$$\therefore r(X, Y) = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = 0$$

$$\rho = 0.$$

Note : E(X) and E(XY) are equal to zero, noted not find σ_x&σ_y.

Example:16

If $y = 2x - 3$ and $y = 5x + 7$ are the two regression lines, find the mean values of x and y. Find the correlation co-efficient between x and y. Find an estimate of x when $y = 1$.

Solution: Given $y = 2x - 3$ ----------(1)

$y = 5x + 7$ ----------(2)

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines.

$\bar{y} = 2\bar{x} - 3$ --------(3)

$\bar{y} = 5\bar{x} + 7$ ---------(4)

Subtracting the equations (3) and (4), we have

$$3\bar{x} = -10 \Rightarrow \bar{x} = \frac{-10}{3}$$

$$\bar{y} = 2\left(\frac{-10}{3}\right) - 3 = \frac{-29}{3}$$

Therefore mean values are $\bar{x} = \frac{-10}{3}$ and $\bar{y} = \frac{-29}{3}$.

Let us suppose that equation (1) is the line of regression of y on x and equation (2) is the equation of the line of regression of x on y, we have

$$(1) \Rightarrow y = 2x - 3$$

$$b_{yx} = 2$$

$$(2) \Rightarrow 5x = y - 7$$

$$x = \frac{1}{5}y - \frac{7}{5}$$

$$b_{xy} = \frac{1}{5}$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{1}{5} \times 2} = \pm 0.63$$

Since both the regression co-efficients are positive, r must be positive.

Correlation co-efficient = $r = 0.63$

Substituting $y = 1$ in (2), we have

$$5x = 1 - 7 = -6$$

$$x = -\frac{6}{5}.$$

3.6 Transforms of Two-Dimensional Random Variable

Formula:

$$f_U(u) = \int_{-\infty}^{\infty} f_{u,v}(u,v)\,dv$$

$$\& \quad f_V(u) = \int_{-\infty}^{\infty} f_{u,v}(u,v)\,du$$

$$f_{UV}(u,V) = f_{XY}(x,y)\left|\frac{\partial(x,y)}{\partial(u,v)}\right|$$

Example : 1

If the joint pdf of (X, Y) is given by $f_{xy}(x, y) = x+y$, $0 \le x, y \le 1$, find the pdf of $\cup = XY$.

Solution

Given $f_{xy}(x, y) = x + y$

Given $U = XY$

Let $V = Y$

$$x = \frac{u}{v} \;\&\; y = V$$

$$\frac{\partial x}{\partial u} = \frac{1}{V}, \frac{\partial x}{\partial v} = \frac{-u}{V^2}; \frac{\partial y}{\partial u} = 0; \frac{\partial y}{\partial v} = 1$$

$$\therefore J = \left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{V} & \frac{-u}{V^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{v} - 1 = \frac{1}{v}$$

$$\Rightarrow |J| = \frac{1}{V}$$

The joint p.d.f. (u, v) is given by

$$f_{uv}(u,v) = f_{xy}(x,y)\,|J|$$

$$= (x+y)\frac{1}{|v|}$$

$$= \frac{1}{V}\left(\frac{u}{v} + u\right)$$

The range of V :

Since $0 \le y \le 1$, we have $0 \le V \le 1$ $(\therefore V = y)$

The range of u :

Given $0 \le x \le 1$

$$\Rightarrow 0 \le \frac{u}{v} \le 1$$

$\Rightarrow \quad 0 \le u \le v$

Hence the p.d.f. of (u, v) is given by

$$f_{uv}(u,v) = \frac{1}{v}\left(\frac{u}{v}+v\right),\ 0 \le u \le v,\ 0 \le v \le 1$$

Now

$$f_U(u) = \int_{-\infty}^{\infty} f_{u,v}(u,v)\,dv$$

$$= \int_{u}^{1} f_{u,v}(u,v)\,dv$$

$$= \int_{u}^{1}\left(\frac{u}{v^2}+1\right)dv$$

$$= \left[v + u.\frac{v^{-1}}{-1}\right]_u^1$$

$\therefore f_u(u) = 2(1-u),\ 0 < u < 1$

p.d.f of (u, v) p.d.f of u = XY

$f_{uv}(u,v) = \frac{1}{v}\left(\frac{u}{v}+v\right)$ $f_u(u) = 2(1-u),\ 0 < u < 1$

$0 \le u \le v,\ 0 \le v \le 1$

Example:18

The pdf of (X,Y) is given by

$f_{XY}(x,y) = e^{-(x+y)}\ ;x \ge 0\ ,\ y \ge 0\ ,$

$U = \frac{X+Y}{2}.$

find the pdf

Solution: Given $f_{XY}(x,y) = e^{-(x+y)}\ ;x \ge 0\ ,\ y \ge 0$

Introduce the auxiliary random variable $V = Y$

$$U = \frac{X+Y}{2} \qquad V = Y$$

$$u = \frac{x+y}{2} \qquad v = y$$

$$2u = x+y \qquad y = v$$

$$2u = x+v$$

$$x = 2u - v$$

$$\frac{\partial x}{\partial u} = 2 \qquad \frac{\partial y}{\partial u} = 0$$

$$\frac{\partial x}{\partial v} = -1 \qquad \frac{\partial y}{\partial v} = 1$$

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 2 & -1 \\ 0 & 1 \end{vmatrix} = 2 - 0 = 2$$

$$|J| = |2| = 2$$

Therefore, the joint density function of UV is given by

$$f_{UV}(u,v) = |J|\, f_{XY}(x,y) \qquad x \geq 0 \ , \ y \geq 0$$

$$= 2e^{-(x+y)} \qquad 2u - v \geq 0 \ , \ v \geq 0$$

$$= 2e^{-(2u - v + v)} \qquad 2u \geq v \qquad , v \geq 0$$

$$= 2e^{-2u} \qquad 0 \leq v \leq 2u$$

The pdf of U is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u,v)dv = \int_0^{2u} 2e^{-2u}\, dv = 2e^{-2u} \int_0^{2u} dv = 2e^{-2u}\,[v]_0^{2u}$$

$$= 2e^{-2u}(2u - 0) = 4ue^{-2u} \quad , u \geq 0 .$$

3.7 Central Limit Theorem

The central limit theorem (CLT) is one of the most important results in probability theory. Ittates that, under certain conditions, the sum of a large number of random variables is approximately normal. Here, we state a version of the CLT that applies to independent and identically distributed (i.i.d.) random variables. Suppose that X1, X2 , ... , Xn are i.i.d. random variables with expected values E(Xi)=μ<∞ and variance Var(Xi)=σ^2<∞. Then as we saw above, the sample mean $\bar{X} = \frac{X_1+X_2+X_3+\cdots+X_n}{n}$ has mean E$\bar{X}$= μ and variance Var($\bar{X}$) $= \frac{\sigma^2}{n}$.

Thus, the normalized random variable Zn = $(\bar{X} - \mu)/(\sigma/\sqrt{n}) = \frac{X_1+X_2+X_3+\cdots+X_n}{\sigma\sqrt{n}}$

Has mean EZn = 0 and variance(Zn)=1. The central limit theorem states that the CDF of Zn converges to the standard normal CDF.

The Central Limit Theorem (CLT)

Let $X_1,X_2,...,X_n$ be i.i.d. random variables with expected value $EX_i = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \qquad \text{for all } x \in \mathbb{R},$$

where $\Phi(x)$ is the standard normal CDF.

Example:19

A bank teller serves customers standing in the queue one by one. Suppose that the service time Xi for customer ii has mean EXi=2 (minutes) and Var(Xi)=1. We assume that service times for different bank customers are independent. Let Y be the total time the bank teller spends serving 50 customers. Find P(90<Y<110).

Solution:

$$Y = X_1 + X_2 + \ldots + X_n,$$

where $n = 50$, $EX_i = \mu = 2$, and $\mathrm{Var}(X_i) = \sigma^2 = 1$. Thus, we can write

$$\begin{aligned} P(90 < Y \leq 110) &= P\left(\frac{90 - n\mu}{\sqrt{n}\sigma} < \frac{Y - n\mu}{\sqrt{n}\sigma} < \frac{110 - n\mu}{\sqrt{n}\sigma}\right) \\ &= P\left(\frac{90 - 100}{\sqrt{50}} < \frac{Y - n\mu}{\sqrt{n}\sigma} < \frac{110 - 100}{\sqrt{50}}\right) \\ &= P\left(-\sqrt{2} < \frac{Y - n\mu}{\sqrt{n}\sigma} < \sqrt{2}\right). \end{aligned}$$

By the CLT, $\frac{Y-n\mu}{\sqrt{n}\sigma}$ is approximately standard normal, so we can write

$$\begin{aligned} P(90 < Y \leq 110) &\approx \Phi(\sqrt{2}) - \Phi(-\sqrt{2}) \\ &= 0.8427 \end{aligned}$$

Example:20

In a communication system each data packet consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1. It is assumed bit errors occur independently. Find the probability that there are more than 120 errors in a certain data packet.

Solution:

Let us define X_i as the indicator random variable for the ith bit in the packet. That is, $X_i = 1$ if the ith bit is received in error, and $X_i = 0$ otherwise. Then the X_i's are i.i.d. and $X_i \sim Bernoulli(p = 0.1)$. If Y is the total number of bit errors in the packet, we have

$$Y = X_1 + X_2 + \ldots + X_n.$$

Since $X_i \sim Bernoulli(p = 0.1)$, we have

$$EX_i = \mu = p = 0.1, \qquad \mathrm{Var}(X_i) = \sigma^2 = p(1-p) = 0.09$$

Using the CLT, we have

$$\begin{aligned} P(Y > 120) &= P\left(\frac{Y - n\mu}{\sqrt{n}\sigma} > \frac{120 - n\mu}{\sqrt{n}\sigma}\right) \\ &= P\left(\frac{Y - n\mu}{\sqrt{n}\sigma} > \frac{120 - 100}{\sqrt{90}}\right) \\ &\approx 1 - \Phi\left(\frac{20}{\sqrt{90}}\right) \\ &= 0.0175 \end{aligned}$$

Summary

In this chapter we have learn

- Joint distribution in both discrete and continuous two-dimensional random variables
- Marginal, Conditional Distribution of two-dimensional random variables
- Covariance and Correlation between two variables.
- Study of the relationship between the variable
- Angle between two lines
- Transformation of random variables
- Central limit theorem

3.10 Keywords

JOINT PROBABILITY DENSITY FUNCTION:

$$P\left[x - \frac{dx}{2} \le X \le x + \frac{dx}{2},\ y - \frac{dy}{2} \le Y \le y + \frac{dy}{2}\right] = f(x,y)\,dx\,dy$$

MARGINAL PROBABILITY FUNCTION[M.P.F]:

$P_{i*} = \sum_i P_{ij}$; $P_{*j} = \sum_j P_{ij}$ (discrete)

The m.p.f of X and Y is given by

$$f_x(x) = f(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{(continuous)}$$

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

CONDITIONAL PROBABILITY FUNCTION[CPF]:

The c.p.m.f of X and Y is given by

$$P(X = x_i \ / Y = y_j) = \frac{P(X=x_i \cap Y=y_j)}{P(Y=y_j)} = \frac{P_{ij}}{P_{*j}} \text{ (discrete)}$$

$$P(Y = y_j \ / X = x_i) = \frac{P(Y=y_j \cap X=x_i)}{P(X=x_i)} = \frac{P_{ij}}{P_{i*}}$$

The c.p.d.f of X and Y is given by

$$f\left(\frac{x}{y}\right) = \frac{f(x,y)}{f(y)} \ ; f\left(\frac{y}{x}\right) = \frac{f(x,y)}{f(x)}$$

COVARIANCE:

$$Cov(X,Y) = E(XY) - E(X) * E(Y)$$

$$Cov(X+a, Y+b) = Cov(X,Y)$$

$$Cov(aX, bY) = abCov(X,Y)$$

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Lines of Regression: The line of Regression of y on x is given by

$y - \bar{y} = b_{yx}(x - \bar{x})$ where $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

The line of Regression of x on y is given by

$x - \bar{x} = b_{xy}(y - \bar{y})$ where $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

TRANSFORMATION OF RANDOM VARIABLES:

Then the joint PDF of (U,V) is given by

$$f_{UV}(u,v) = |J| f_{XY}(x,y) \text{ where } J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

SELF-ASSESSMENT QUESTIONS

1. The joint p.d.f of the RV(X, Y) is given by $f(x,y) = \ Kxye^{-(x^2+y^2)}, x > 0, y > 0$. Find the value of K and prove also that X & Y are independent.

2. If the joint p.d.f of a two-dimensional r.v (X,Y) is given by,

   $f(x,y) = \begin{cases} x^2 + \frac{xy}{3}, 0 < x < 1; 0 < y < 2 \\ 0, otherwise \end{cases}$

   Find (i) $P\left(x > \frac{1}{2}\right)$ (ii)$P(Y < X)$ and (iii) $P\left(P\left(y < \frac{1}{2} / X < \frac{1}{2}\right)\right)$ Check whether the conditional density functions are valid.

3. Given $f_{xy}(x,y) = \ Cx(x-y), 0 < x < 2, -x < y < x \ and \ 0 \ elsewhere.$

   (i)Evaluate C (iii) Find $f_X(x)$ (iii)$f_{Y/X}(\frac{y}{x})$ (iv) $f_Y(y)$

4. If the joint p.d.f of a 2D RV(X,Y) is given by

$f(x,y) = \begin{cases} K(6-x-y), 0 < x < 2, 2 < y < 4 \\ 0, otherwise \end{cases}$

Find (i) the value of K (ii) $P(X < 1, Y < 3)$ (iii) $P(X + Y < 3)$(iv)$P(X < 1/\ Y < 3)$
(v) $P(X < 1 \cap\ Y < 3)$ or $P(X < 1, Y < 3)$

5.

If the joint density function of the two RVs X and Y be $f(x,y) = \begin{cases} e^{-(x+y)}, x \geq 0, y \geq 0 \\ 0, otherwise \end{cases}$

Find (i) $P(X < 1)$ and ii) $P(X + Y\ < 1)$

6. Given the joint pdf of (X,Y) as $f(x,y) = \begin{cases} 8xy, 0 < x < y < 1 \\ 0, otherwise \end{cases}$ Find the marginal and conditional pdfs of X and Y . Are X and Y are independent?

7. The joint pdf of X and Y is given by $f(x,y) = e^{-(x+y)}, x > 0, y > 0$, find the pdf of $U = \frac{X+Y}{2}$

8. If X and Y are independent r.v each normally distributed with mean zero and variance σ^2, find the density functions of $R = \sqrt{X^2 + Y^2}$ and $\theta = \tan^{-1}\left(\frac{Y}{X}\right)$

9. If X and Y are independent RVs uniformly distributed in (0,1) obtain the distribution of X+Y,XY.

10. If X and Y are independent r.vs with joint pdf $f(x,y) = 3e^{-(x+3y)}, x \geq 0, y \geq 0$, find the p.d.f of $Z = \frac{X}{Y}$

11. If X and Y are independent r.vs with pdf'S $e^{-x}, x \geq 0, e^{-y} y \geq 0$ respectively, find the density function of $U = \frac{X}{X+Y}$ and V = X+Y. Are U and V independent?

12. The marks obtained by 10 students in Mathematics and Statistics are given below. Find the correlation coefficient between the two subjects.

Marks in Mathematics:	75	30	60	80	53	35	15	40	38
48									
Marks in Statistics:	85	45	54	91	58	63	35	43	45
44									

13. Compute the correlation coefficient between X and y using the following data:

X:	1	3	5	7	8	10
Y:	8	12	15	17	18	20

14. Suppose X and Y are r.v's having the joint probability function given by

$f(x,y) = \begin{cases} 4xye^{-(x^2+y^2)}, & x,y \geq 0, \\ 0, & otherwise \end{cases}$

Obtain the p.d.f of $U = \sqrt{X^2 + Y^2}$

15. Find the constant k such that $f(x,y) = \begin{cases} k(x+1)e^{-y}, & 0 < x < 1, y > 0 \\ 0 & otherwise \end{cases}$ is a joint p.d.f of the continuous R.V (X,Y).Are X and Y are independent R.Vs ? Explain.

16. If the joint density function of the two RVs X and Y be $f(x,y) = \begin{cases} e^{-(x+y)}, x \geq 0, y \geq 0 \\ 0, otherwise \end{cases}$

Find the p.d.f of the R.V $U = \frac{X}{Y}$

18. Let the joint p.d.f of r.v (X,Y) be given as $f(x,y) = \begin{cases} Cxy^2, & 0 \le x \le y \le 1 \\ 0 & ,otherwise \end{cases}$ Determine (1) Value of C (2) the marginal p.d.fs of X and Y (3) the conditional $f\left(\frac{x}{y}\right)$ of X given Y=y.

19. The joint p.m.f of two r.v X and Y is given by $P_{XY}(x,y) = \begin{cases} K(2x+y) & x = 1,2\,; y = 1,2 \\ 0 & otherwise \end{cases}$ where K is a constant (1) Find K (2) Find the marginal PMFs of x and Y

20. Two random variables X and Y are related as Y=4X+9. Find the correlation coefficient between X and Y.

21. If the density function is defined by $f(x,y) = \frac{y}{(1+x)^4} e^{\frac{-y}{1+x}} \quad x \ge 0, y \ge 0$ then obtain the regression equation of Y on X for the distribution.

22. Assume that the random variable S_n is the sum of 48 independent experimental values of the random variable X whose PDF is given by $f_X(x) = \begin{cases} \frac{1}{3} & 1 \le x \le 4 \\ 0 & otherwise \end{cases}$ Find the probability that S_n lies in the range $108 \le S_n \le 126$

FURTHER READINGS

1. Devore, J.L, Probability and Statistics for Engineering and Sciences, Cengage Learning, 8th Edition, New Delhi, 2014.
2. Miller and M. Miller, Mathematical Statistics, Pearson Education Inc., Asia 7th Edition, New Delhi, 2011.
3. Richard Johnson, Miller and Freund's Probability and Statistics for Engineer, Prentice Hall of India Private Ltd., 8th Edition, New Delhi, 2011.

UNIT IV TESTING OF HYPOTHESIS

CONTENTS

Learning Objectives

In this chapter a student has to learn the

- Testing of Hypothesis
- Different types of sampling
- To testing significant difference problem involving Large samples
- To testing significant difference problem involving Small samples
- To testing significant difference using t-test, Z-test, F-test, Chi Square test
- Tests for independence of attributes and goodness of fit.

Learning Outcomes

Upon completion of this Unit, students are able to demonstrate a good understanding of:

- Difference between Large and small sample
- Handling sampling distribution using t-test, z-test
- Identify significance difference between sample and population.

Overview

In this Unit, you are going to study about the sampling. Significant difference between sample and population means in both small and large sample. Application of t-test, z-test when various parameters given. To learning application of F-test, when variance given. Testing for independence of attributes and goodness of fit.

4.1 Introduction

In real life, we work with data that are affected by randomness, and we need to extract information and draw conclusions from the data. The randomness might come from a variety of sources.

The randomness might come from a variety of sources. Here are two examples of such situations:

Suppose that we would like to predict the outcome of an election. Since we cannot poll the entire population, we will choose a random sample from the population and ask them who they plan to vote for. In this experiment, the randomness comes from the sampling. Note also that if our poll is conducted one month before the election, another source of randomness is that people might change their opinions during the one month period.

In a wireless communication system, a message is transmitted from a transmitter to a receiver. However, the receiver receives a corrupted version (a noisy version) of the transmitted signal. The receiver needs to extract the original message from the received noisy version. Here, the randomness comes from the noise.

Examples like these are abundant. Dealing with such situations is the subject of the field of statistical inference.

Statistical inference is a collection of methods that deal with drawing conclusions from data that are prone to random variation.

Random Sampling

When collecting data, we often make several observations on a random variable. For example, suppose that our goal is to investigate the height distribution of people in a well-defined population (i.e., adults between 25 and 50 in a certain country). To do this, we define random variables X1, X2, X3, ..., Xn as follows: We choose a random sample of size n with replacement from the population and let Xi be the height of the ith chosen person. More specifically,

We chose a person uniformly at random from the population and let X1 be the height of that person. Here, every person in the population has the same chance of being chosen.

To determine the value of X2, again we choose a person uniformly (and independently from the first person) at random and let X2 be the height of that person. Again, every person in the population has the same chance of being chosen.

In general, Xi is the height of the ith person that is chosen uniformly and independently from the population.

Clearly, we use our knowledge of probability theory when we work on statistical inference problems. However, the big addition here is that we need to work with real data. The probability problems that we have seen in this book so far were clearly defined and the probability models were given to us.

For example, you might have seen a problem like this:

Let X be a normal random variable with mean $\mu=100$ and variance $\sigma^2=15$.
Find the probability that $X>110$.

In real life, we might not know the distribution of X, so we need to collect data, and from the data we should conclude whether X has a normal distribution or not. Now, suppose that we can use the central limit theorem to argue that X is normally distributed. Even in that case, we need to collect data to be able estimate μ and σ.

Here is a general setup for a statistical inference problem: There is an unknown quantity that we would like to estimate. We get some data. From the data, we estimate the desired quantity.

Point Estimation

Here, we assume that θ is an unknown parameter to be estimated. For example, θ might be the expected value of a random variable, $\theta=EX$. The important assumption here is that θ is a fixed (non-random) quantity. To estimate θ, we need to collect some data. Specifically, we get a random sample X1,X2, X3, ...,Xn such that Xi's have the same distribution as X. To estimate θ, we define a point estimator $\hat{\theta}$ that is a function of the random sample

$\hat{\theta}=h(X1,X2,\cdots,Xn)$. For example, if $\theta=EX$, we may choose $\hat{\theta}$ to be the sample mean

$$\hat{\theta}=\bar{X}=\frac{X1+X2+X3+\ldots\ldots+Xn}{n}$$

Now, suppose that we would like to estimate the variance of a distribution σ^2. Assuming $0<\sigma^2<\infty$, by definition

$\sigma^{\wedge}2=\frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2$.

Interval Estimation (Confidence Intervals)

Let X1,X2, X3, ...,Xn be a random sample from a distribution with a parameter θ that is to be estimated. An interval estimator with confidence level 1−α consists

$\hat{\theta}$l(X1,X2,⋯,Xn) and $\hat{\theta}$h(X1,X2,⋯,Xn) such that P($\hat{\theta}$l ≤θ and$\hat{\theta}$h≥θ)≥1−α, for every possible value of θ. Equivalently, we say that [$\hat{\theta}$l, $\hat{\theta}$h] is a (1−α)100% confidence interval for θ.

Hypothesis

What is Hypothesis Testing?
Hypothesis is usually considered as the principal instrument in research. The main goal in many research studies is to check whether the data collected support certain statements or predictions. A statistical hypothesis is an assertion or conjecture concerning one or more populations. Test of hypothesis is a process of testing of the significance regarding the parameters of the population on the basis of sample drawn from it. Thus, it is also termed as "Test of Significance'.

Points to be considered while formulating Hypothesis

- Hypothesis should be clear and precise.
- Hypothesis should be capable of being tested.
- Hypothesis should state relationship between variables.
- Hypothesis should be limited in scope and must be specific.
- Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned.
- Hypothesis should be amenable to testing within a reasonable time.
- Hypothesis must explain empirical reference.

Types of Hypothesis:
There are two types of hypothesis, i.e., Research Hypothesis and Statistical Hypothesis
1. Research Hypothesis: A research hypothesis is a tentative solution for the problem being investigated. It is the supposition that motivates the researcher to accomplish future course of action. In research, the researcher determines whether or not their
supposition can be supported through scientific investigation.
2. Statistical Hypothesis: Statistical hypothesis is a statement about the population which we want to verify on the basis of sample taken from population. Statistical hypothesis is stated in such a way that they may be evaluated by appropriate statistical techniques.

4.2 Types of Statistical Hypotheses
There are two types of statistical hypotheses:
1. Null Hypothesis (H0) – A statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
2. Alternative Hypothesis (H1 or Ha) –

A statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters. Alternative hypothesis is created in a negative meaning of the null hypothesis.

4.3 BASIC CONCEPTS CONCERNING TESTING OF HYPOTHESES

1. **The level of significance:** This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that H0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H0 is true. In other words, the 5 per cent level of significance means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesis when it (H0) happens to be true. Thus, the significance level is the maximum value of the probability of rejecting H0 when it is true and is usually determined in advance before testing the hypothesis.
2. **Decision rule or Test of Hypothesis:** A decision rule is a procedure that the researcher uses to decide whether to accept or reject the null hypothesis. The decision rule is a statement that tells under what circumstances to reject the null hypothesis. The decision rule is based on specific values of the test statistic (e.g., reject H0 if Calculated value > table value at the same level of significance)
3. **Types of Error:** In the context of testing of hypotheses, there are basically two types of errors we can make.

   Type 1 error: Type I error, also known as a "false positive": the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. Plainly speaking, it occurs when we are observing a difference when in truth there is none (or more specifically - no statistically significant difference). So the probability of making a type I error in a test with rejection region R is $P(R\backslash H_0$ is true), often denoted by α.

   Type II error: Type II error, also known as a "false negative": the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. Plainly speaking, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is $1-P(R\backslash H_a$ is true) often denoted as β

   In a tabular form the said two errors can be presented as follows:

Particulars	Decision	
	Accept H_0	**Reject H_0**
H_0 (True)	Correct Decision	Type I error (α error)
H_0 (False)	Type II error (β error)	Correct decision

4. **One- tailed and Two-tailed Tests:** A test of statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one tailed test.

   A test of statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test.

   Numerical Steps in Testing of Hypothesis

   1. Establish the null hypothesis and alternative hypothesis.
   2. set up a suitable significance level e.g.at 1%, 5%, 10% level of significance etc.
   3. Determine a suitable test tool like t, Z, F, Chi Square, ANOVA etc.
   4. Calculate the value of test statistic using any of test tools
   5. Compare this calculated value with table value

6. Draw conclusions

If calculated value < Table value then null hypothesis is accepted

If calculated value > Table value then null hypothesis is rejected.

Tools available for testing Hypothesis

1. Large Sample tests
2. Small Sample tests

4.4 Large Sample tests

When the sample size is n≥30, then apply large sample tests.

4.4.1 Test for the significant difference between sample mean and population mean

Testing of significance for single mean

To find significant difference between mean of sample and population

test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ When population S.D. is known

test statistic is Z = $Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ When population S.D. is not known

Example 1: A sample of 100 people during the past year showed an average life span of 71.8 years. If the standard deviation of the population is 8.9 years, test whether the mean life span today is greater than 70 years.

Solution : Given sample size n = 100 sample mean $\bar{x}$ = 71.8 years

Population Mean μ = 70 years

Population s.d σ = 8.9

We want to test whether μ > 70

Null hypothesis H_0 : μ = 70 (i.e., there is no significant difference between sample mean and population mean)

Alternative hypothesis H_1 : μ > 70 (i.e., Right tailed test)

Under H_0 , the test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$$= \frac{71.8 - 70}{\frac{8.9}{\sqrt{100}}} = \frac{1.8 \times 10}{8.9} = 2.02$$

$\therefore$ Z = 2.02

The table value of Z at 5% level is 1.645

Inference :

Since Z > 1.645, H_0 is rejected at 5% level of significance.

$\therefore$ μ > 70 is acceptable (i.e., the mean life span today is greater than 70 years)

Example 2: A sample of 100 students is taken from a large population. The mean height of the students in this sample is 160 cm. Can it be reasonably regarded that this sample is from a population of mean 165 cm and the SD 10cm? Also find the 95% fiducial limits for the mean.

Solution: Given sample size n = 100 sample mean $\bar{x} = 160$

Population Mean μ = 165

Population s.d $\sigma = 10$

We want to test the difference between sample mean and population mean.

Null hypothesis H_0 : μ = 165(i.e., there is no significant difference between sample mean and population mean)

Alternative hypothesis H_1 : $\mu \neq 165$ (i.e., there is a significant difference between the sample mean and population mean)

$\therefore$ Two tailed test

Under H_0 , the test statistic is $Z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

$$= \frac{160 - 165}{\dfrac{10}{\sqrt{100}}} = -5$$

$\therefore \ |Z| = 5$

The table value of $|Z|$ at 5% level is 1.96

Inference :

Since $|Z| > 1.96$, H_0 is rejected at 5% level of significance.

$\therefore$ $\mu \neq 165$ (i.e., there is a significant difference between the sample mean and population mean)

Example 3: A sample of 900 items has mean 3.4 cms and standard deviation 2.61cms. Can the sample be regarded as drawn from a population with mean 3.25 cms at 5% level of significance?

Solution: Given sample size n = 900 sample mean $\bar{x} = 3.4$

Population Mean μ = 3.25

Sample standard s = 2.61 σ is not known.

We have to test whether the sample be drawn from population with mean 3.25 cm.

Null hypothesis H_0 : μ = 3.25

Alternative hypothesis H_1 : $\mu \neq 3.25$ (i.e., Two tailed test)

Under H_0 , the test statistic is $Z = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{n}}}$

$$= \frac{3.4 - 3.25}{\dfrac{2.61}{\sqrt{900}}} = \frac{0.15 \times 30}{2.61} = 1.72$$

$\therefore \ |Z| = 1.72$

The table value of $|Z|$ at 5% level is 1.96

Inference :

Since $|Z| < 1.96$, H_0 is accepted at 5% level of significance.

i.e., the sample can be regarded as drawn from a population with mean 3.25cm.

Example 4: The mean lifetime of a sample of 100 light tubes produced by a company is found to be 1580 hours with standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the tubes produced by the company is 1600 hours.

Solution: Given n = 100 $\bar{x} = 1580$

$\mu = 1600$ s = 90 σ is not known.

Null hypothesis H_0: $\mu = 1600$ (i.e., there is no significant difference between sample mean and population mean)

Alternative hypothesis H_1: $\mu \neq 1600$ (i.e., Two tailed test)

Under H_0, the test statistic is $Z = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{n}}}$

$$= \frac{1580 - 1600}{\frac{90}{\sqrt{100}}} = -2.22$$

$\therefore |Z| = 2.22$

The table value of $|Z|$ at 5% level is 1.96

Inference:

Since $|Z| > 1.96$, H_0 is rejected at 5% level of significance.

i.e., the mean lifetime of the tubes produced by the company may not be 1600 hours.

Example 5: The average number of defective articles produced per day in a certain factory is claimed to be less than all the factories. The average of all the factories is 30.5. A random sample of 100 days production showed the mean defective as 28.8 and standard deviation 6.35. Is the average less than 30.5 for all the factories?

Solution : Given n = 100 $\bar{x} = 28.8$

$\mu = 30.5$ s = 6.35 σ is not known.

We want to test the average article produced is less than 30.5

Null hypothesis H_0 : $\mu = 30.5$

Alternative hypothesis H_1 : $\mu < 30.5$ (i.e., left tailed test)

Under H_0 , the test statistic is $Z = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{n}}}$

$$= \frac{28.8 - 30.5}{\frac{6.35}{\sqrt{100}}} = -2.68$$

$\therefore$ Z = 2.68

The table value of Z for left tailed test at 1% level is – 2.33 and at 5% level is – 1.645

Inference :

Since Z < - 1.645 and Z < - 2.33 , H_0 is rejected at 5% and 1% level of significance.

$\therefore$ $\mu < 30.5$ for all factories.

4.4.2 Test for the significant difference between two means

Testing of significance for difference of means

test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ When population S.D. is known

test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ When population S.D. is not known

Example 6: Two types of new cars produced in our country are tested for petrol mileage. One group consisting of 36 cars averaged 14 km per litre. While the other group consisting of 72 cars averaged 12.5 km per litre. i) what test statistic is appropriate, if $\sigma_1^2 = 1.5, \sigma_2^2 = 2$?

ii)Test, whether there exists a significant difference in the petrol consumption of these two type cars.

Solution:

Sample I	Sample II
$n_1 =$ 36	$n_2 =$ 72
$\bar{x}_1 =$ 14 km per litre	$\bar{x}_2 =$ 12.5 km per litre
Population variance $\sigma_1^2 = 1.5$	Population variance $\sigma_2^2 = 2$

The appropriate test statistic to be used is the test of difference between the means

Null hypothesis $H_0 : \mu_1 = \mu_2$ (i.e., there is no difference between the mean)

$H_1 : \mu_1 \neq \mu_2$

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

$$Z = \frac{14 - 12.5}{\sqrt{\dfrac{1.5}{36} + \dfrac{2}{72}}} = \frac{1.5}{\sqrt{0.0694}} = 5.69$$

$\therefore$ $|Z| = 5.69$

The table value of $|Z|$ at 5% level is 1.96

Inference :

Since $|Z| > 1.96$, H_0 is rejected at 5% level of significance (i.e., the difference of means is highly significance.

Example 7: A sample of heights of 6400 Englishmen has a mean of 170 cms and a standard deviation of 6.4 cms, while a sample of heights of 1600 Australians has a mean of 172 cm and standard deviation of 6.3cm. Do the data indicate that the Australians are on the average taller than the Englishmen.

Solution : Englishmen (Sample I) Australians (Sample II)

$n_1 = 6400$ $n_2 = 1600$

$\bar{x}_1 = 170$ cm $\bar{x}_2 = 172$ cm

$s_1 = 6.4$ cm $s_2 = 6.3$ cm

Let μ_1 be the mean height of the population of Englishmen and μ_2 be the mean height of the population of Australians. We want to test whether Australians are taller than Englishmen.

Null hypothesis H_0 : $\mu_1 = \mu_2$

Alternative hypothesis H_1 : $\mu_1 < \mu_2$ (i.e., left tailed test)

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

$$Z = \frac{170 - 172}{\sqrt{\dfrac{6.4^2}{6400} + \dfrac{6.3^2}{1600}}} = \frac{-2}{\sqrt{0.031}} = -11.32$$

$\therefore$ Z = - 11.32

The table value of Z for left tailed test at 1% level is – 2.33 and at 5% level is – 1.645

Inference :

Since Z < - 1.645 and Z < - 2.33 , H_0 is rejected at 5% and 1% level of significance.

$\therefore$ Australians are taller than Englishmen.

Example 8: Intelligence test on two groups of boys and girls gave the following results.

	Mean	S.D	Sample Size
Girls	75	15	150
Boys	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls?

Solution : Girls (Sample I) Boys (Sample II)

$n_1 = 150$ $n_2 = 250$

$\bar{x}_1 = 75$ $\bar{x}_2 = 70$

$s_1 = 15$ $s_2 = 20$

We want to test whether the mean scores obtained by boys an girls equivalent.

H_0 : $\mu_1 = \mu_2$ (There is no significant difference between the mean scores obtained by boys and girls.

H_1 : $\mu_1 \neq \mu_2$ (i.e., two tailed test)

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

$$Z = \frac{75 - 70}{\sqrt{\dfrac{15^2}{150} + \dfrac{20^2}{250}}} = \frac{5}{\sqrt{3.1}} = 2.84$$

$$\therefore \ |z| = 2.84$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z > 2.58 , H_0 is rejected at 5% and 1% level of significance. (i.e., the difference is highly significant).

Example 9: The mean height of two samples of 1000 and 2000 members are respectively 67.5 and 68 inches. Can they be regarded as drawn from the same population with standard deviation 2.5 inches.

Solution :

Sample I	Sample II
$n_1 = 1000$	$n_2 = 2000$
$\bar{x}_1 = 67.5$ inches	$\bar{x}_2 = 68$ inches

Population standard deviation $\sigma = 2.5$

We want to test whether the significant difference between the two sample means.

H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$ (i.e., two tailed test and σ is known)

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$$Z = \frac{67.5 - 68}{2.5\sqrt{\dfrac{1}{1000} + \dfrac{1}{2000}}} = -5.12$$

$$\therefore \ |z| = 5.12$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z > 2.58 , H_0 is rejected at 5% and 1% level of significance. (i.e., the difference is highly significant).

Hence the two samples cannot be regarded as drawn from the same population with σ = 2.5.

Example 10: The mean produce of wheat from a sample of 100 fields comes to 200kg per acre and another sample of 150 fields gives a mean of 220 kg per acre. Assuming the S.D of the yield at 11kg for the universe, test if there is a significant difference between the means of the samples.

Solution :

Sample I	Sample II
$n_1 = 100$	$n_2 = 150$
$\bar{x}_1 = 200$	$\bar{x}_2 = 220$

Population standard deviation $\sigma = 11$

We want to test whether the significant difference between the two sample means.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$ (i.e., two tailed test and σ is known)

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$$Z = \frac{200 - 220}{11\sqrt{\dfrac{1}{100} + \dfrac{1}{150}}} = -14.08$$

$$\therefore |z| = 14.08$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since $Z > 1.96$ and $Z > 2.58$, H_0 is rejected at 5% and 1% level of significance. (i.e., the difference is highly significant).

Example 11: Two random sample of sizes 400 and 500 have mean 10.9 and 11.5 respectively. Can the samples be regarded as drawn from the same population with variance 25?

Solution :

Sample I	Sample II
$n_1 = 400$	$n_2 = 500$
$\bar{x}_1 = 10.9$	$\bar{x}_2 = 11.5$

Population standard deviation $\sigma^2 = 25 \Rightarrow \sigma = 5$

We want to test whether the two samples are drawn from a population with variance 25

$H_0 : \mu_1 = \mu_2$ (i.e., there is nosignificant difference between the two population means)

$H_1 : \mu_1 \neq \mu_2$ (i.e., two tailed test and σ is known)

Under H_0 the test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$$Z = \frac{10.9 - 11.5}{5\sqrt{\dfrac{1}{400} + \dfrac{1}{500}}} = -1.78$$

$$\therefore |z| = 1.78$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z < 1.96 and Z < 2.58 , H_0 is accepted at 5% and 1% level of significance. (i.e., the samples can be regarded as drawn from the same population with variance 25).

4.4.3 Test for the significant difference between the standard deviation of two large samples

To find significant difference between two sample S. D.

test statistic is $Z = \dfrac{s_1 - s_2}{\sigma\sqrt{\dfrac{1}{2n_1} + \dfrac{1}{2n_2}}}$ When population S.D. is known

test statistic is $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$ When population S.D. is not known

Example 12: The standard deviation of a population is 3. Samples of 1000 and 500 are drawn from it and their standard deviation are respectively 2.6 and 2. Is the difference between standard deviations significant?

Solution :

Sample I	Sample II
$n_1 = 1000$	$n_2 = 500$
$s_1 = 2.6$	$s_2 = 2$

Population standard deviation $\sigma = 3$

We want to test whether the two samples are drawn from the same population with standard deviation.

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \neq \sigma_2$ (i.e., two tailed test) and σ_1, σ_2 is unknown

Under H_0 the test statistic is $Z = \dfrac{s_1 - s_2}{\sigma\sqrt{\dfrac{1}{2n_1} + \dfrac{1}{2n_2}}}$

$$Z = \frac{2.6 - 2}{3\sqrt{\dfrac{1}{2000} + \dfrac{1}{1000}}} = 5.16$$

$$\therefore |z| = 5.16$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z > 2.58 , H_0 is rejected at 5% and 1% level of significance.

Example 13: Two samples of 100 and 80 bulbs of a factory are selected at random from the same production batch. The mean and standard deviation of the first batch are 540 hours and 30 hours and that of the second batch are 552 hours and 28 hours. Do you think the difference between the two standard deviations is significant?

Solution :

Sample I	Sample II
$n_1 = 100$	$n_2 = 80$
$s_1 = 30$	$s_2 = 28$

$H_0 : \sigma_1 = \sigma_2$ (i.e., there is no significant difference between the two S.D)

$H_1 : \sigma_1 \neq \sigma_2$ (i.e., two tailed test and σ is unknown)

Under H_0 the test statistic is $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$

$$Z = \frac{30 - 28}{\sqrt{\dfrac{30^2}{2000} + \dfrac{28^2}{160}}} = 0.86$$

$$\therefore |z| = 0.86$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since $Z < 1.96$ and $Z < 2.58$, H_0 is accepted at 5% and 1% level of significance. (i.e., the difference between the two standard deviations is not significant)

Example 14: A sample of heights of 6400 Englishmen have a standard deviation of 6.4 cm while a sample of heights of 1600 Australians have standard deviation 6.3 cm. Is this difference significant?

Solution :

Sample I	Sample II
$n_1 = 6400$	$n_2 = 1600$
$s_1 = 6.4$	$s_2 = 6.3$

$H_0 : \sigma_1 = \sigma_2$ (i.e., there is no significant difference between the two S.D)

$H_1 : \sigma_1 \neq \sigma_2$ (i.e., two tailed test) and σ_1, σ_2 is unknown

Under H_0 the test statistic is $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$

$$Z = \frac{6.4 - 6.3}{\sqrt{\dfrac{6.4^2}{2(6400)} + \dfrac{6.3^2}{2(1600)}}} = 0.8$$

$$\therefore |z| = 0.8$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z < 1.96 and Z < 2.58 , H_0 is accepted at 5% and 1% level of significance. (i.e., the difference between the two standard deviations is not significant)

Example 15: From the table discuss the difference between the standard deviations is significant or not.

	Group A	Group B
No. of items	50	60
S.D	4	4.2

Solution : Sample I Sample II

$n_1 = 50$ $n_2 = 60$

$s_1 = 4$ $s_2 = 4.2$

$H_0 : \sigma_1 = \sigma_2$ (i.e., there is no significant difference between the two S.D)

$H_1 : \sigma_1 \neq \sigma_2$ (i.e., two tailed test) and σ_1, σ_2 is unknown

Under H_0 the test statistic is $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}}$

$$Z = \frac{4 - 4.2}{\sqrt{\dfrac{4^2}{2(50)} + \dfrac{4.2^2}{2(60)}}} = -0.36$$

$\therefore |z| = 0.36$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z < 1.96 and Z < 2.58 , H_0 is accepted at 5% and 1% level of significance. (i.e., the difference between the two standard deviations is not significant)

Example 16: Two samples of sizes 1000 and 2000 are drawn from a normal population with standard deviation 40. The standard deviation of the first sample was found to 42 and that of the second was found to be 44. Is the difference significant?

Solution : Sample I Sample II

$n_1 = 1000$ $n_2 = 2000$

$s_1 = 42$ $s_2 = 44$

Population standard deviation $\sigma = 40$

We want to test whether the two samples are drawn from the same population with standard deviation.

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \neq \sigma_2$ (i.e., two tailed test) and σ_1, σ_2 is unknown

Under H_0 the test statistic is $Z = \dfrac{s_1 - s_2}{\sigma\sqrt{\dfrac{1}{2n_1} + \dfrac{1}{2n_2}}}$

$$Z = \frac{42 - 44}{40\sqrt{\dfrac{1}{2(1000)} + \dfrac{1}{2(2000)}}} = -1.8$$

$$\therefore \ |z| = 1.8$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since $Z < 1.96$ and $Z < 2.58$, H_0 is accepted at 5% and 1% level of significance. (i.e., the difference between the two standard deviations is not significant)

4.4.4 Test for the significant difference between sample proportion and population proportion

To find significant difference between proportion of sample and population test statistic is $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

Example 17: A coin is tossed 900 times and head appears 490 times. Does this support the hypothesis that the coin is unbiased?

Solution: Given n = 900, Let p be the proportion of heads when a coin is tossed 900 times. Since head appears 490 times.

$$p = \frac{490}{900} = 0.54$$

P = proportion of getting head in the population

$= \dfrac{1}{2}$, if the coin is unbiased.

$$Q = 1 - P = \frac{1}{2}$$

We want to test whether the coin is unbiased or not.

$H_0 : P = \dfrac{1}{2}$ (i.e., the coin is unbiased)

$H_1 : P \neq \dfrac{1}{2}$ (i.e., Two tailed test)

Under H_0 the test statistic is $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

$$Z = \frac{0.544 - 0.5}{\sqrt{\frac{1}{2}.\frac{1}{2}.\frac{1}{900}}} = 2.64$$

$$\therefore \ |z| = 2.64$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z > 2.58 , H_0 is rejected at 5% and 1% level of significance.

$\therefore$ The date does not support the hypothesis that the coin is unbiased.

Example 18: A quality control engineer suspects that the proportion of defective units among certain manufactured items has increased from the set limit of 0.01. To test the claim, he randomly selected 100 of these items and found that the proportion of defective units in the sample was 0.02. Test the engineer's hypothesis at the 0.05 level of significance. **Solution** : Given Sample size n = 100

Let p = Proportion of defective units in the sample =0.02

P = Proportion of defective units in the population = 0.01

Q = 1 – P = 1 – 0.01 = 0.99

H_0 : P = 0.01

H_1 : P > 0.01(i.e., Right tailed test)

Under H_0 the test statistic is $Z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$

$$Z = \frac{0.02-0.01}{\sqrt{\dfrac{0.01(0.99)}{100}}} = 1.01$$

$\therefore Z = 1.01$

The table value of Z at 5% level is 1.645

Inference :

Since Z < 1.645 , H_0 is accepted at 5% level of significance.

$\therefore$ The claim P = 0.01 is true.

Example 19: In a sample of 1000 people in Mumbai, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 5% level of significance?

Solution: Given n = 1000 and No. of rice eaters = 540

$$p = \frac{540}{1000} = 0.54$$

H_0 : P = 0.5(i.e., both rice and wheat are equally popular in the state)

H_1 : P $\neq$ 0.5(i.e., Two tailed test)

Under H_0 the test statistic is $Z = \dfrac{p-P}{\sqrt{\dfrac{PQ}{n}}}$

$$Z = \frac{0.54-0.5}{\sqrt{\dfrac{0.5(0.5)}{100}}} = 2.53$$

$\therefore |z| = 2.53$

The table value of $|z|$ 5% level is 1.96

Inference :

Since Z > 1.96 , H_0 is rejected at 5% level of significance.

Example 20: 40 people were attacked by a disease and only 36 survived. Will your reject the hypothesis that the survival rate, if attacked by this disease, is 85% in favour of the hypothesis that it is more at 5% level of significance.

Solution : Given Sample size n = 40 and No. of survivors = 36

Let p = Proportion of defective units in the sample =0.85

$$p = \frac{36}{40} = 0.9$$

Q = 1 – P = 1 – 0.85 = 0.15

H_0 : P = 0.85 (i.e., the proportion of persons survived after attack by disease in the lot is 85%)

H_1 : P > 0.85(i.e., Right tailed test)

Under H_0 the test statistic is $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$$Z = \frac{0.9 - 0.85}{\sqrt{\frac{0.85(0.15)}{40}}} = 0.89$$

$\therefore Z = 0.89$

The table value of Z at 5% level is 1.645

Inference :

Since Z < 1.645 , H_0 is accepted at 5% level of significance.

Example 21: In a sample of 400 parts produced by a factory, the number of defective parts was found to be 30. The company however claims that only 5% of their products is defective. Is the claim tenable?

Solution : Given Sample size n = 400 and No. of defectives in the sample = 30

Let p = Proportion of defective in the sample

$$p = \frac{30}{400} = 0.08$$

P = Proportion of defective parts in the population = 5% =0.05

Q = 1 – P = 1 – 0.05 = 0.95

We want to test whether the proportion in the population in P = 5% or not.

H_0 : P = 0.05

H_1 : P ≠ 0.05(i.e., Two tailed test)

Under H_0 the test statistic is $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$$Z = = \frac{0.08 - 0.05}{\sqrt{\frac{0.05(0.95)}{400}}} = 2.75$$

$\therefore |z| = 2.75$

The table value of $|z|$ 5% level is 1.96

Inference :

Since Z > 1.96 , H_0 is rejected at 5% level of significance.

$\therefore$ The claim is not tenable.

Example 22: A die is thrown 9000 times and throw of 3 or 4 is observed 3240 times. Show that the die cannot be regarded as an unbiased one and find the limits between which the probability of a throw of 3 or 4 lies.

Solution: Given n = 9000

Let success be defined as throwing 3 or 4.

Let p = Proportion of defective in the sample $= \frac{3240}{9000} = 0.36$

Let P = Probability of success = $\frac{1}{3}$

H_0 : P = $\frac{1}{3}$ (i.e., the die is unbiased)

H_1 : P $\neq \frac{1}{3}$ (i.e., Right tailed test)

Under H_0 the test statistic is $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$$Z = \frac{0.36 - 0.3333}{\sqrt{\frac{0.3333(0.6667)}{9000}}} = 5.37$$

$\therefore |z| = 5.37$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z > 2.58 , H_0 is rejected at 5% and 1% level of significance.

$\therefore$ The date does not support the hypothesis that the coin is unbiased.

4.4.5 Test for the significant difference between two proportions in two samples

To find significant difference between two sample proportions

test statistic is $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where Q = 1 – P

we estimate P = $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ When P is not given

Example 23 : Random samples of 400 men and 600 women were asked whether they would like to have a fly-over near their residence 200 men and 325 women were in favor of it. Test the equality of proportion of men and women in the proposal?

Solution: Given $n_1 = 400$ $n_2 = 600$

p_1 = Proportion of men in favor of the proposal = $\frac{200}{400} = 0.5$

p_2 = Proportion of women in favor of the proposal = $\frac{325}{600} = 0.54$

Since P is not given, we estimate it as P = $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$= \frac{400(0.5) + 600(0.54)}{400 + 600} = 0.52$$

$$Q = 1 - P = 1 - 0.53 = 0.47$$

H_0: $p_1 = p_2$ (i.e., the proportions of men and women in favor of the proposal are the same)

H_1 : $p_1 \neq p_2$ (i.e., Two tailed test)

Under H_0 the test statistic is Z = $\frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$Z = \frac{0.5 - 0.54}{\sqrt{0.53(0.47)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.3$$

$\therefore |z| = 1.3$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z < 1.96 and Z < 2.58 , H_0 is accepted at 5% and 1% level of significance. (i.e., men and women are equally favorable for fly over near their residence).

Example 24: In a random sample of 1000 people from city A, 400 are found to be consumers of wheat. In a sample of 800 from city B, 400 are found to be consumers of wheat. Does this data give a significant difference between the two cities as far as the proportion of wheat consumers is concerned?

Solution: Given $n_1 = 1000$ $n_2 = 800$

p_1 = Proportion of wheat consumers in city A = $\frac{400}{1000} = 0.4$

p_2 = Proportion of wheat consumers in cityB = $\frac{400}{800} = 0.5$

Since P is not given, we estimate it as P = $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$= \frac{1000(0.4) + 800(0.5)}{1000 + 800} = 0.44$$

$$Q = 1 - P = 1 - 0.44 = 0.56$$

H_0: $p_1 = p_2$ (i.e., there is no significant difference in the proportions of wheat consumers in the two cities)

H_1 : $p_1 \neq p_2$ (i.e., Two tailed test)

Under H_0 the test statistic is Z = $\frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$Z = \frac{0.4 - 0.5}{\sqrt{0.44(0.56)\left(\frac{1}{1000} + \frac{1}{800}\right)}} = -4.3$$

$$\therefore \ |z| = 4.3$$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z >2.58 , H_0 is rejected at 5% and 1% level of significance. (i.e., there is a significant difference between the two cities as far as the proportion of wheat consumers is concerned)

Example 25: In a year there are 956 births in a town A of which 52.5% were male, while in towns A and B combined, this proportion in a total of 1406 births was 0.496. Is there any significant difference in the proportion of male births in the two towns?

Solution: Given $n_1 = 956$ and $n_1 + n_2 = 1406$

$$n_2 = 1406 - 956 = 450$$

p_1 = Proportion of male births in town A = 0.525

p_2 = Proportion of male births in town B (unknown)

P = Proportion of males in the combined sample = 0.496 (given)

i.e., $0.496 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \Rightarrow 0.496 = \frac{986(0.525) + 450(p_2)}{956 + 450}$

$$p_2 = 0.434$$

H_0: $p_1 = p_2$ (i.e., there is no significant difference in the proportions of male births in the two towns)

H_1 : $p_1 \neq p_2$ (i.e., Two tailed test)

Under H_0 the test statistic is Z = $\frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$Z = \frac{0.525 - 0.434}{\sqrt{0.496(0.504)\left(\frac{1}{956} + \frac{1}{450}\right)}} = 3.37$$

$\therefore \ |z| = 3.37$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z > 1.96 and Z >2.58 , H_0 is rejected at 5% and 1% level of significance. (i.e., there is a significant difference in the proportion of male births in the two towns)

Example 26: In a random sample of 100 men taken from village A, 60 were found to be consuming alcohol. In another sample of 200 men taken from village B, 100 were found to be consuming alcohol. Do the two villages differ significantly in respect to the proportion of men who consume alcohol?

Solution: Given $n_1 = 100 \quad n_2 = 200$

p_1 = Proportion of men consuming alcohol in village A = $\frac{60}{100} = 0.6$

p_2 = Proportion of men consuming alcohol in village B = $\frac{100}{200} = 0.5$

Since P is not given, we estimate it as $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$= \frac{100(0.6) + 200(0.5)}{100 + 200} = 0.53$$

$Q = 1 - P = 1 - 0.53 = 0.47$

H_0: $p_1 = p_2$

H_1 : $p_1 \neq p_2$ (i.e., Two tailed test)

Under H_0 the test statistic is $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$Z = \frac{0.6 - 0.5}{\sqrt{0.53(0.47)\left(\frac{1}{100} + \frac{1}{200}\right)}} = 1.64$$

$\therefore \ |z| = 1.64$

The table value of $|z|$ 5% level is 1.96 and at 1% level is 2.58

Inference :

Since Z < 1.96 and Z < 2.58 , H_0 is accepted at 5% and 1% level of significance.

Example 27: Before an increase in excise duty on tea, 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase in duty, 800 people were tea drinkers

in a sample of 1200 people. State whether there is a significant decrease in the consumption of tea after the increase in excise duty?

Solution: Given $n_1 = 1000$ $n_2 = 1200$

p_1 = Proportion of tea drinkers before increase in excise duty

$= \frac{800}{1000} = 0.8$

p_2 = Proportion of tea drinkers after increase in excise duty

$= \frac{800}{1200} = 0.67$

Since P is not given, we estimate it as $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$= \frac{1000(0.8) + 1200(0.67)}{1000 + 1200} = 0.73$

$Q = 1 - P = 1 - 0.73 = 0.27$

H_0: $p_1 = p_2$ (i.e., there is no significant difference in the consumption of tea before and after the increase in excise duty)

H_1 : $p_1 > p_2$ (i.e., Right tailed test)

Under H_0 the test statistic is $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$Z = \frac{0.8 - 0.67}{\sqrt{0.73(0.27)\left(\frac{1}{1000} + \frac{1}{1200}\right)}} = 6.8$

$\therefore$ Z = 6.8

The table value of Z at 5% level is 1.645

Inference :

Since Z > 1.645, H_0 is rejected at 5% level of significance.

level of significance. (i.e., there is no significant difference between the people consuming tea before and after increase in excise duty)

Example 28: A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved?

Solution: Given $n_1 = 400$ $n_2 = 300$

p_1 = Proportion of defective articles produced by the machine before overhauling $= \frac{20}{400} =$ 0.05

p_2 = Proportion of defective articles produced by the machine after overhauling $= \frac{10}{300} = 0.03$

Since P is not given, we estimate it as $P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$$= \frac{400(0.05)+300(0.03)}{400+300} = 0.04$$

Q = 1 – P = 1 – 0.04 = 0.96

H_0: $p_1 = p_2$

H_1 : $p_1 > p_2$ (i.e., Right tailed test)

Under H_0 the test statistic is $Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

$$Z = \frac{0.05-0.03}{\sqrt{0.04(0.95)\left(\frac{1}{400}+\frac{1}{300}\right)}} = 1.3$$

$\therefore$ Z = 1.1

The table value of Z at 5% level is 1.645

Inference :

Since Z < 1.645, H_0 is accepted at 5% level of significance.

level of significance.

$\therefore$ The machine has not improved after overhauling

4.5 Exact Sampling Distributions (t, F, χ^2)

Small Sample Tests: When the sample size is n<30 then apply small sample tests.

4.5.1 Test the significant difference between sample mean and population mean

test statistic is $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$

Example 29: Ten oil tins are taken from an automatic filling machine. The mean weight of the tins is 15.8kg and S.D 0.5kg. Does the sample mean differ significantly from the intended weight 16 kg?

Solution: Given n = 10, $\bar{x}$ = 15.8kg, s = 0.5kg, μ = 16kg

H_0: μ = 16 kg (i.e., the sample mean weight is not different than the intended weight)

H_1 : $\mu \neq 16$ kg (i.e., two tailed test)

Under H_0 the test statistic is $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}} = \frac{15.8-16}{\frac{0.5}{\sqrt{10}}} = -1.2$ $\therefore |t| = 1.2$

No. of degree of freedom $\nu = n-1 = 10-1 = 9$

For $\nu = 9$ degree of freedom, the table value of t at 5% level is $t_{0.05} = 2.26$

Inference :

Since calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The difference between sample mean weight and the intended weight is not significant.

Example 30: The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

Solution: Given n = 10, $\mu = 64$

H_0: $\mu = 64$ (i.e., the average height is equal to 64 inches)

H_1 : $\mu > 64$ (i.e., Right tailed test)

$$\bar{x} = \frac{\sum x_i}{n} = \frac{660}{10} = 66 \qquad S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{9}(90) = 10 \qquad \therefore s = \sqrt{10}$$

Under H_0 the test statistic is $t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \dfrac{66-64}{\frac{\sqrt{10}}{\sqrt{10}}} = 2$ $\qquad \therefore\ t = 2$

No. of degree of freedom $\nu = n - 1 = 10 - 1 = 9$

For $\nu = 9$ degree of freedom, the table value for Right tailed test at 5% level is $t_{0.05} = 1.833$

Inference :

Since calculated value of t > the table value of t, H_0 is rejected at 5% level of significance.

$\therefore$ The average height is greater then 64 inches.

Example 31: Eleven articles produced by a factory were chosen at random and their weights were found to be (in kgs) 63, 63, 66, 67, 68, 69, 70, 70, 71, 71 and 71 respectively. In the light of the above data, Can we assume that the mean weight of articles produced by the factory is 66kgs.

Solution: Given n = 11, $\mu = 66$

H_0: $\mu = 66$ (i.e., the data are consistent with the assumption of mean weight of 11 articles produced by the factory)

H_1 : $\mu \neq 66$ kg (i.e., two tailed test)

$$\bar{x} = \frac{\sum x_i}{n} = \frac{749}{11} = 68.09 \qquad S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = 9.12 \qquad \therefore s = 3.02$$

Under H_0 the test statistic is $t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \dfrac{68.09 - 66}{\frac{3.02}{\sqrt{11}}} = 2.3$ $\qquad \therefore\ |t| = 2.3$

No. of degree of freedom $\nu = n - 1 = 11 - 1 = 10$

For $\nu = 10$ degree of freedom, the table value of t at 5% level is $t_{0.05} = 2.23$

Inference :

Since calculated value of t > the table value of t, H_0 is rejected at 5% level of significance.

$\therefore$ The mean weight of the articles produced by factory cannot be assumed to be 66kgs.

Example 32: A random sample of 10boys had the following IQ's : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean IQ of 100?

Solution: Given n = 10, $\mu = 100$

H_0: $\mu = 100$ (i.e., the data are consistent with the assumption of mean IQ of 100 in population)

H_1 : $\mu \neq 100$ (i.e., two tailed test)

$$\bar{x} = \frac{\sum x_i}{n} = \frac{972}{10} = 97.2 \qquad S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{9}(1833.6) = 203.73$$

$$\therefore s = 14.27$$

Under H_0 the test statistic is t = $\dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}} = \dfrac{97.2 - 100}{\dfrac{14.27}{\sqrt{10}}} = -0.62$ $\therefore$ $|t|$

= 0.62

No. of degree of freedom $\nu = n - 1 = 10 - 1 = 9$

For $\nu = 9$ degree of freedom, the table value of t at 5% level is $t_{0.05}$ = 2.262

Inference :

Since calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The mean IQ of the population can be 100.

4.5.2 Test of significance of the difference between the means of the two samples

test statistic is t = $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

Example 33: Two independent sample from normal populations with equal variance gave the following.

Sample	Sample Size	Mean	S.D
1	16	23.4	2.5
2	12	24.9	2.8

Is the difference between the means significant.

Solution: Given $n_1 = 16$ $n_2 = 12$

$\bar{x}_1 = 23.4$ $\bar{x}_2 = 24.9$

$s_1 = 2.5$ $s_2 = 2.8$

We want to test whether the significance of difference between the mean of the two samples

H_0: $\mu_1 = \mu_2$ (i.e., there is no significant difference between means)

$H_1: \mu_1 \neq \mu_2$ (i.e., two tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{23.4 - 24.9}{\sqrt{\dfrac{16(2.5)^2 + 12(2.8)^2}{16+12-2}\left(\dfrac{1}{16} + \dfrac{1}{12}\right)}} = -1.37$$

$$|t| = 1.37$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 16 + 12 - 2 = 26$

For $\nu = 26$ the table value of t at 5% level is $t_{0.05} = 2.06$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The difference between the means is not significant.

Example 34: Two random samples gave the following results

Sample	Size	Sample Mean	Sum of the squares of deviations from the mean
1	10	15	90
2	12	14	108

Examine whether the samples come from the same normal population.

Solution : Solution: Given $n_1 = 10 \qquad n_2 = 12$

$$\bar{x}_1 = 15 \qquad \bar{x}_2 = 14$$

$$\sum (x_1 - \bar{x}_1)^2 = 90 \qquad \sum (x_2 - \bar{x}_2)^2 = 108$$

$$s_1^2 = \frac{1}{n_1}\sum (x_1 - \bar{x}_1)^2 = \frac{90}{10} = 9 \quad s_2^2 = \frac{1}{n_2}\sum (x_2 - \bar{x}_2)^2 = \frac{108}{12} = 9$$

We want to test whether the two samples come from the same normal population.

H_0: $\mu_1 = \mu_2$ (i.e., there is no significant difference between means)

$H_1: \mu_1 \neq \mu_2$ (i.e., two tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{15 - 14}{\sqrt{\dfrac{16(9) + 12(9)}{10+12-2}\left(\dfrac{1}{10} + \dfrac{1}{12}\right)}} = 0.74$$

$$|t| = 0.74$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$

For $\nu = 20$ the table value of t at 5% level is $t_{0.05} = 2.086$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

Example 35: Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results.

HorseA	28	30	32	33	33	29	34
HorseB	29	30	30	24	27	29	

Test whether Horse A is running faster than B at 5%level.

Solution: Given $n_1 = 7 \qquad n_2 = 6$

Now we shall find the sample means $\bar{x}_1, \bar{x}_2$ and their variances s_1^2, s_2^2 from the data

Horse A				Horse B		
x_1	$d_1 = x_1 - 33$	d_1^2		x_2	$d_2 = x_2 - 30$	d_2^2
28	-5	25		29	-1	1
30	-3	9		30	0	0
32	-1	1		30	0	0
33	0	0		24	-6	36
33	0	0		27	-3	9
29	-4	16		29	-1	1
34	1	1				
	-12	52			-11	47

$$\bar{x}_1 = 33 + \frac{\sum d_1}{n_1} \qquad \bar{x}_2 = 30 + \frac{\sum d_2}{n_2}$$

$$= 33 - \frac{12}{7} = 31.29 \qquad = 30 - \frac{11}{6} = 28.17$$

$$s_1^2 = \frac{\sum d_1^2}{n_1} - \left(\frac{\sum d_1}{n_1}\right)^2 \qquad s_2^2 = \frac{\sum d_2^2}{n_2} - \left(\frac{\sum d_2}{n_2}\right)^2$$

$$= \frac{52}{7} - \left(\frac{-12}{7}\right)^2 = 4.5 \qquad = \frac{47}{6} - \left(\frac{-11}{6}\right)^2 = 4.47$$

H_0: $\mu_1 = \mu_2$ (i.e., there is no significant difference between means)

H_1 : $\mu_1 > \mu_2$ (i.e., Horse B runs faster than Hourse A and right tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{31.29 - 28.17}{\sqrt{\dfrac{7(4.5) + 6(4.48)}{7 + 6 - 2}\left(\dfrac{1}{7} + \dfrac{1}{6}\right)}} = 2.43$$

$$t = 2.43$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 7 + 6 - 2 = 11$

For $\nu = 11$ degree of freedom, the table value of t at 5% level for right tailed test is $t_{0.05} = 1.796$

Inference :Since the calculated value of t > the table value of t, H_0 is rejected at 5% level of significance.

$\therefore$ Horse B runs faster than Horse A.

Example 36: The following are the number of sales which a sample of 9 sales people of industrial chemicals in Gujarat and a sample of 6 sales people of industrial chemicals in Mumbai made over a certain fixed period of time.

Gujarat	59	68	44	71	63	46	69	54	48
Mumbai	50	36	62	52	70	41			

Solution: Given $n_1 = 9 \qquad n_2 = 6$

Now we shall find the sample means $\bar{x}_1, \bar{x}_2$ and their variances s_1^2, s_2^2 from the data

x_1	$d_1 = x_1 - 54$	d_1^2		x_2	$d_2 = x_2 - 50$	d_2^2
59	5	25		50	0	0
68	14	196		36	-14	196
44	-10	100		62	12	144
71	17	289		52	2	4
63	9	81		70	20	400
46	-8	64		41	-9	
69	15	225				
54	0	0				
48	-6	36				
	36	1016			11	825

$$\bar{x}_1 = 54 + \frac{\sum d_1}{n_1} \qquad \bar{x}_2 = 50 + \frac{\sum d_2}{n_2}$$

$$= 54 + \frac{36}{9} = 58 \qquad = 50 + \frac{11}{6} = 51.83$$

$$s_1^2 = \frac{\sum d_1^2}{n_1} - \left(\frac{\sum d_1}{n_1}\right)^2 \qquad s_2^2 = \frac{\sum d_2^2}{n_2} - \left(\frac{\sum d_2}{n_2}\right)^2$$

$$= \frac{1016}{9} - \left(\frac{36}{9}\right)^2 = 96.89 \qquad = \frac{825}{6} - \left(\frac{11}{6}\right)^2 = 134.14$$

H_0: $\mu_1 = \mu_2$ (i.e., there is no significant difference between means)

H_1 : $\mu_1 \neq \mu_2$ (i.e., two tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{58-51.83}{\sqrt{\frac{9(96.89)+6(134.14)}{9+6-2}(\frac{1}{9}+\frac{1}{6})}} = 1.03$$

$$|t| = 1.03$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 9+6-2 = 13$

For $\nu = 13$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 2.16$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

Example 37: A group of 10 rats fed on a diet A and another group of 8 rats fed on a different diet B, recorded the following increase in weight. Does it show the superiority of diet A over diet B.

DietA	5	6	8	1	12	4	3	9	6	10
DietB	2	3	6	8	1	10	2	8		

Solution: Given $n_1 = 10 \qquad n_2 = 8$

Now we shall find the sample means $\bar{x}_1, \bar{x}_2$ and their variances s_1^2, s_2^2 from the data

x_1	d_1^2		x_2	d_2^2
5	25		2	4
6	36		3	9
8	64		6	36
1	1		8	64
12	144		1	1
4	16		10	100
3	9		2	4
9	81		8	64
6	36			
10	100			
64	512		40	282

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{64}{10} = 6.4 \qquad \bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{40}{8} = 5$$

$$s_1^2 = \frac{\sum x_1^2}{n_1} - \left(\frac{\sum x_1}{n_1}\right)^2 \qquad s_2^2 = \frac{\sum x_2^2}{n_2} - \left(\frac{\sum x_2}{n_2}\right)^2$$

$$= \frac{512}{10} - \left(\frac{64}{10}\right)^2 = 10.24 \qquad = \frac{282}{8} - \left(\frac{40}{8}\right)^2 = 10.25$$

H_0: $\mu_1 = \mu_2$ (i.e., no significant difference between the two diets)

H_1 : $\mu_1 > \mu_2$ (i.e., diet A superior to diet B) (right tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{6.4 - 5}{\sqrt{\dfrac{10(10.24) + 8(10.25)}{10 + 8 - 2}\left(\dfrac{1}{10} + \dfrac{1}{8}\right)}} = 0.869$$

$$t = 0.869$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 10 + 8 + -2 = 16$

For $\nu = 16$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 1.176$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The difference is not significant.

Example 38: The table below represent the values of protein content from cow's milk and buffalo's milk at a certain level. Examine if these differences are significant.

Cow's Milk	1.82	2.02	1.88	1.61	1.81	1.54
Buffalo's Milk	2	1.83	1.86	2.03	2.19	1.88

Solution: Given $n_1 = 6 \qquad n_2 = 6$

Now we shall find the sample means $\bar{x}_1, \bar{x}_2$ and their variances s_1^2, s_2^2 from the data

Cow's Milk Buffalo's Milk

x_1	x_1^2		x_2	x_2^2
1.82	3.3124		2	1.0040
2.02	4.0804		1.83	3.3489
1.88	3.5344		1.86	3.4596
1.61	2.5921		2.03	4.1209
1.81	3.2761		2.19	4.7961
1.54	2.3716		1.88	3.5344
10.68	19.1670		11.79	23.2599

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{10.68}{6} = 1.78$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{11.79}{6} = 1.965$$

$$s_1^2 = \frac{\sum x_1^2}{n_1} - \left(\frac{\sum x_1}{n_1}\right)^2$$

$$= \frac{19.167}{6} - \left(\frac{10.68}{6}\right)^2 = 0.0261$$

$$s_2^2 = \frac{\sum x_2^2}{n_2} - \left(\frac{\sum x_2}{n_2}\right)^2 = \frac{23.2599}{6} - \left(\frac{11.79}{6}\right)^2 = 0.0154$$

H_0: $\mu_1 = \mu_2$ (i.e., no significant difference between the means)

H_1 : $\mu_1 \neq \mu_2$ (i.e., two tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$$= \frac{1.78 - 1.965}{\sqrt{\dfrac{6(0.0261) + 6(0.0154)}{6 + 6 - 2}\left(\dfrac{1}{6} + \dfrac{1}{6}\right)}} = -2.03$$

$$|t| = 2.03$$

No. of degree of freedom $\nu = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$

For $\nu = 10$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 2.26$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The difference between the mean protein contents of cow's milk and buffalo's milk is not significant.

4.6 t-Test for paired Observations

test statistic is $t = \dfrac{\bar{d}}{\dfrac{s}{\sqrt{n-1}}}$ Where $d = x_1 - x_2 or x_2 - x_1$

$$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 :$$

Example 39: The following are the average weekly loss of working hours due to accidents in 10 industrial plants before and after an introduction of a safety program was put into operation.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Use 0.05 level of significance to test whether the safety program is effective.

Solution: Given $n = 10$

Let μ be difference between the means of the populations before and after the safety program.

H_0: $\mu = 0$ (i.e., The safety program is not effective)

H_1 : $\mu > 0$ (i.e., right tailed test)

Under H_0 the test statistic is $t = \dfrac{\bar{d}}{\dfrac{s}{\sqrt{n-1}}}$ Where $d = x_1 - x_2 or x_2 - x_1$

x_1	x_2	$d = x_2 - x_1$	d^2
45	36	-9	81
73	60	-13	169
46	44	-2	4
124	119	-5	25
33	35	2	4
57	51	-6	36
83	77	-6	36
34	29	-5	25
26	24	-2	4
17	11	-6	36
		-52	420

$\sum d = -52$, $\sum d^2 = 420$ and $\bar{d} = \frac{\sum d}{n} = \frac{-52}{10} = -5.2$

$$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{420}{10} - (-5.2)^2 = 14.96 \Rightarrow s = 3.868$$

$$t = \frac{-5.2}{\frac{3.868}{\sqrt{10-1}}} = -4.03$$

$$|t| = 4.03$$

No. of degree of freedom $\nu = n - 1 = 10 - 1 = 9$

For $\nu = 9$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 1.833$

Inference :

Since the calculated value of t > the table value of t, H_0 is rejected at 5% level of significance.

$\therefore$ The safety program is effective.

Example 40: A certain stimulus administered to each of 12 patients resulted in the following change in blood pressure (B.P) 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.

Solution : We are given the increments in blood pressure $d = x_2 - x_1$

We want to test whether the change in blood pressure is significant or not.

Let μ be difference between the mean of the change is blood pressure after the stimulus.

H_0: $\mu = 0$ (i.e., change in blood pressure is not significant)

H_1 : $\mu > 0$ (i.e., right tailed test)

d	5	2	8	-1	3	0	-2	1	5	0	4	6

d^2	25	4	64	1	9	0	4	1	25	0	16	36

Under H_0 the test statistic is $t = \dfrac{\bar{d}}{\dfrac{s}{\sqrt{n-1}}}$

$\sum d = 31, \sum d^2 = 185$ and $\bar{d} = \dfrac{\sum d}{n} = \dfrac{31}{12} = 2.583$

$$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{420}{10} - (2.583)^2 = 8.745 \Rightarrow s = 2.957$$

$$t = \frac{2.583}{\dfrac{2.957}{\sqrt{12-1}}} = 2.9$$

$$|t| = 2.9$$

No. of degree of freedom $\nu = n - 1 = 12 - 1 = 11$

For $\nu = 11$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 1.796$

Inference :

Since the calculated value of t > the table value of t, H_0 is rejected at 5% level of significance.

$\therefore$ The stimulous generally increases the blood pressure.

Example 41: Eleven school boys were given a test in Mathematics. They were given one month tution and the second test was held at the end of it. Do the marks provide evidence that the students were benefitted by the extra coaching.

Marks I	23	20	19	21	18	20	18	17	23	16	19
Marks II	24	19	22	18	20	22	20	20	23	20	18

Solution : Given n = 11

Let μ be difference between marks of all the students before and after the coaching .

We want to test whether the training effective or not.

H_0: $\mu = 0$ (i.e.,The students are not benefitted by extra coaching)

H_1 : $\mu > 0$ (i.e., right tailed test)

x_1	x_2	$d = x_2 - x_1$	d^2
23	24	1	1
20	19	-1	1
19	22	3	9
21	18	-3	9
18	20	2	4
20	22	2	4

18	20	2	4
17	20	3	9
23	23	0	0
16	20	4	16
19	18	-1	1
		12	58

$\sum d = 12$, $\sum d^2 = 58$ and $\bar{d} = \frac{\sum d}{n} = \frac{12}{11} = 1.09$

$$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{420}{10} - (1.09)^2 = 4.085 \Rightarrow s = 2.02$$

Under H_0 the test statistic is $t = \frac{\bar{d}}{\frac{s}{\sqrt{n-1}}}$ Where $d = x_1 - x_2 or x_2 - x_1$

$$t = \frac{1.09}{\frac{2.02}{\sqrt{11-1}}} = 1.7$$

$$|t| = 1.7$$

No. of degree of freedom $\nu = n - 1 = 11 - 1 = 10$

For $\nu = 10$ degree of freedom, the table value of t at 5% level of significance is $t_{0.05} = 1.812$

Inference :

Since the calculated value of t < the table value of t, H_0 is accepted at 5% level of significance.

$\therefore$ The students are not benefitted by extra coaching.

4.7 F-Distribution

Test for the equality of variances of two populations

test statistic is $F = \frac{S_1^2}{S_2^2}$

$$s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1} \qquad s_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2}$$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} \qquad S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

Example 42: Two random samples gave the following results

Sample	1	2
Sample Size	10	12

Sum of squares of deviations from the mean	90	108

Test whether the samples come from the populations with same variance.

Solution: Given $n_1 = 10$ $n_2 = 12$

$\sum (x_i - \bar{x})^2 = 90$ $\sum (y_i - \bar{y})^2 = 108$

Let s_1^2, s_2^2 be the variances of the two samples are σ_1^2, σ_2^2 be the variances of the two populations.

We want to test whether the samples came from two populations with the same variances.

H_0: $\sigma_1^2 = \sigma_2^2$ (i.e., The population variances are equal)

H_1 : $\sigma_1^2 \neq \sigma_2^2$ (i.e., two tailed test)

$$s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1} = \frac{90}{10} = 9 \text{ and } s_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2} = \frac{108}{12} = 9$$

Now $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10(9)}{9} = 10$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{12(9)}{11} = 9.82$

$\therefore S_1^2 > S_2^2$

Under H_0, the

The No. of degrees of freedom is $(\nu_1, \nu_2) = (n_1 - 1, n_2 - 1) = (9,11)$

Degrees of freedom (9,11), the table value of F at 5% level

is $F_{0.05} = 2.9$

Inference :

Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

$\therefore$ The two samples came from two populations with same variance.

Example 43: Two independent samples of sizes 7 and 6 have the following values

Sample A	28	30	32	33	31	29	34
Sample B	29	30	30	24	27	28	

Examine whether the samples have been drawn from normal populations having same variance using 0.05 level of significance.

Solution: Given $n_1 = 7$ $n_2 = 6$

Let s_1^2, s_2^2 be the variances of the two samples are σ_1^2, σ_2^2 be the variances of the two populations.

We want to test whether the samples came from two populations with the same variances.

H_0: $\sigma_1^2 = \sigma_2^2$ (i.e., The population variances are equal)

H_1 : $\sigma_1^2 \neq \sigma_2^2$ (i.e., two tailed test)

Sample A Sample B

x_1	$d_1 = x_1 - 30$	d_1^2		x_2	$d_2 = x_2 - 30$	d_2^2
28	-2	4		29	-1	1
30	0	0		30	0	0
32	2	4		30	0	0
33	3	9		24	-6	36
31	1	1		27	-3	9
29	-1	1		28	-2	4
34	4	16				
	7	35			-12	50

$\sum d_1 = 7, \sum d_2 = -12, \sum d_1^2 = 35, \sum d_2^2 = 50$

$$s_1^2 = \frac{\sum d_1^2}{n_1} - \left(\frac{\sum d_1}{n_1}\right)^2 = \frac{35}{7} - \left(\frac{7}{7}\right)^2 = 4$$

$$s_2^2 = \frac{\sum d_2^2}{n_2} - \left(\frac{\sum d_2}{n_2}\right)^2 = \frac{50}{6} - \left(\frac{-12}{6}\right)^2 = 4.33$$

Now $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{7(4)}{6} = 4.67$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{6(4.33)}{5} = 5.196$

$\therefore S_2^2 > S_1^2$

Under H_0, the test statistic is $F = \frac{S_2^2}{S_1^2} = \frac{5.196}{4.67} = 1.11$

The No.of degrees of freedom is $(\nu_1, \nu_2) = (n_1 - 1, n_2 - 1) = (6,5)$

Degrees of freedom (6,5), the table value of F at 5% level is $F_{0.05} = 4.39$

Inference :

Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

$\therefore$ The two samples came from two populations with same variance.

Example 44: A group of 10 rats fed on diet A and another group of 8 rats fed on diet B, recorded the following increase in weight.

DietA	5	6	8	1	12	4	3	9	6	10
DietB	2	3	6	8	10	1	2	8		

Can you say that the two samples came from the same population?

Solution: Given $n_1 = 10 \qquad n_2 = 8$

Let s_1^2, s_2^2 be the variances of the two samples are σ_1^2, σ_2^2 be the variances of the two populations.

We want to test whether the samples came from two populations with the same variances.

H_0: $\sigma_1^2 = \sigma_2^2$ (i.e., The population variances are equal)

H_1 : $\sigma_1^2 \neq \sigma_2^2$ (i.e., two tailed test)

Diet A Diet B

x_1	x_1^2		x_2	x_1^2
5	25		2	4
6	36		3	9
8	64		6	36
1	1		8	64
12	144		10	100
4	16		1	1
3	9		2	4
9	81		8	64
6	36			
10	100			
64	512		40	282

$\sum x_1 = 64,$ $\sum x_2 = 40, \sum x_1^2 = 512, \sum x_2^2 = 282$

$$s_1^2 = \frac{\sum x_1^2}{n_1} - \left(\frac{\sum x_1}{n_1}\right)^2 = \frac{512}{10} - \left(\frac{64}{10}\right)^2 = 10.24$$

$$s_2^2 = \frac{\sum x_2^2}{n_2} - \left(\frac{\sum x_2}{n_2}\right)^2 = \frac{282}{8} - \left(\frac{40}{8}\right)^2 = 10.25$$

Now $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10(10.24)}{9} = 11.38$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{8(10.25)}{7} = 11.71$ $\therefore S_2^2 > S_1^2$

Under H_0, the test statistic is $F = \frac{S_2^2}{S_1^2} = \frac{11.71}{11.38} = 1.03$

The No. of degrees of freedom is $(\nu_1, \nu_2) = (n_1 - 1, n_2 - 1) = (9, 7)$

Degrees of freedom (9,7), the table value of F at 5% level is $F_{0.05} = 3.29$

Inference :

Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

$\therefore$ The two samples came from two populations with same variance.

Example 45: Two independent samples of sizes 9 and 7 from a normal population had the following values of the variables.

Sample I	18	13	12	15	12	14	16	14	15
SampleII	16	19	13	16	18	13	15		

Do the estimates of the population variance differ significantly at 5% level?

Solution: Given $n_1 = 9$ $n_2 = 7$

Let s_1^2, s_2^2 be the variances of the two samples are σ_1^2, σ_2^2 be the variances of the two populations.

We want to test whether the samples came from two populations with the same variances.

H_0: $\sigma_1^2 = \sigma_2^2$ (i.e., The population variances are equal)

H_1 : $\sigma_1^2 \neq \sigma_2^2$ (i.e., two tailed test)

Sample I Sample II

x_1	x_1^2		x_2	x_1^2

18	324		16	256
13	169		19	361
12	144		13	169
15	225		16	256
12	144		18	324
14	196		13	169
16	256		15	225
14	196			
15	225			
129	1879		110	1760

$\sum x_1 = 129, \sum x_2 = 110, \sum x_1^2 = 1871, \sum x_2^2 = 1760$

$$s_1^2 = \frac{\sum x_1^2}{n_1} - \left(\frac{\sum x_1}{n_1}\right)^2 = \frac{1879}{9} - \left(\frac{129}{9}\right)^2 = 3.33$$

$$s_2^2 = \frac{\sum x_2^2}{n_2} - \left(\frac{\sum x_2}{n_2}\right)^2 = \frac{1760}{7} - \left(\frac{110}{7}\right)^2 = 4.49$$

Now $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9(3.33)}{8} = 3.75$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{7(4.49)}{6} = 5.24$ $\quad \therefore S_2^2 > S_1^2$

Under H_0, the test statistic is $F = \frac{S_2^2}{S_1^2} = \frac{5.24}{3.75} = 1.4$

The No.of degrees of freedom is $(\nu_1, \nu_2) = (n_1 - 1, n_2 - 1) = (8, 6)$

Degrees of freedom (8,6), the table value of F at 5% level is $F_{0.05} = 3.58$

Inference :

Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

$\therefore$ The two samples came from two populations with same variance.

4.8 Chi-square Distribution

χ^2 -Test of Goodness of Fit

test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

Where O – Observed frequency, E- Expected frequency

Example 46: The table below gives the number of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thurs	Fri	Sat

No. of accidents	14	18	12	11	15	14

Solution :

We want to test whether the accidents are uniformly distributed.

H_0 : The accidents are uniformly distributed over the 6 days

H_1 : The accidents are not uniformly distributed.

The expected frequencies for each day $= \frac{84}{6} = 14$

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14	14	0	0	0
18	14	4	16	1.143
12	14	-2	4	0.286
11	14	-3	9	0.643
15	14	1	1	0.071
14	14	0	0	0
				$\chi^2 = 2.143$

Number of degrees of freedom $\nu = n - 1 = 6 - 1 = 5$

For $\nu = 5$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 11.07$

Inference : Since the calculated value of χ^2 < the table value of χ^2, H_0 is accepted at 5% level of significance.

$\therefore$ The accidents are uniformly distributed over the 6 days.

Example 47: In 120 throws of a single die, the following distributions of faces was observed.

Face	1	2	3	4	5	6
Frequency	30	25	18	10	22	15

Can you say that the die is biased.

Solution :

We want to test whether the die is biased or not.

H_0 : The die is unbiased.

H_1 : The die is biased.

The expected frequencies for each day $= \frac{120}{6} = 20$

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$

30	20	10	100	5
25	20	5	25	1.25
18	20	-2	4	0.2
10	20	-10	100	5
22	20	2	4	0.2
15	20	-5	25	1.25
				$\chi^2 = 12.9$

Number of degrees of freedom $\nu = n-1 = 6-1 = 5$

For $\nu = 5$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 11.07$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ The die is biased.

Example 48: A sample analysis of examination results of 500 students was made. It was found that 220 students have failed, 170 have secured a third class, 90 have secured a second class and the rest, a first class. So these figures support the general belief that the above categories are in the ratio 4 : 3 : 2 : 1 respectively?

Solution :

We want to test whether the general examination result is in the ratio 4 : 3 : 2 : 1.

H_0 : The result is in the ratio 4 : 3 : 2 : 1

H_1 : The result is not in the ratio 4 : 3 : 2 : 1.

No. of students failed = 220 Total students = 500

No. of III class students = 170 No. of II class students = 90

No. of I class students = $500-220-170-90 = 20$

The expected frequencies of the 4 classes are

$\frac{4(500)}{10}, \frac{3(500)}{10}, \frac{2(500)}{10}, \frac{1(500)}{10}$ [4+3+2+1=10]

i.e., 200, 150, 100, 50

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
220	200	20	400	2
170	150	20	400	2.667
90	100	-10	100	1
20	50	-30	900	18
				$\chi^2 = 23.67$

Number of degrees of freedom $\nu = n-1 = 4-1 = 3$

For $\nu = 3$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 7.815$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ The result is not in the ratio 4 : 3 : 2 : 1.

Example 49: The following data gives the number of male and female births in 1000 families having five children.

Male Child	0	1	2	3	4	5
Female Child	5	4	3	2	1	0
Frequency	40	300	250	200	130	80

Test whether the given data is consistent with the hypothesis that the binomial law holds with even chance.

Solution : Let us fit a binomial distributions B(X, n, p) to the given data where p be the probability of a male child.

$\therefore P(X = x) = nC_x p^x q^{n-x}, x = 0,1,2,.....n$

We want to test whethere the even chance of getting a male or female child.

H_0 : Male and female births are equally probable.

H_1 : Male and female births are not equally probable.

Probability of male birth $p = \frac{1}{2}$, Probability of female birth $q = \frac{1}{2}$

Here n = 5, $\therefore P(X = x) = 5C_x p^x q^{n-x}, x = 0,1,2,.....5$

Total frequency N = 1000

The expected frequencies of getting 0, 1, 2, 3, 4, 5 success are given by

N X P(X = 0) $= 1000\left[5C_0 (0.5)^0 (0.5)^{5-0}\right] = 31.25 = 31$

N X P(X = 1) $= 1000\left[5C_1 (0.5)^1 (0.5)^{5-1}\right] = 156.25 = 156$

N X P(X = 2) $= 1000\left[5C_2 (0.5)^2 (0.5)^{5-2}\right] = 312.5 = 313$

N X P(X = 3) $= 1000\left[5C_3 (0.5)^3 (0.5)^{5-3}\right] = 312.5 = 313$

N X P(X = 4) $= 1000\left[5C_4 (0.5)^4 (0.5)^{5-4}\right] = 156.3 = 156$

N X P(X = 5) $= 1000\left[5C_5 (0.5)^5 (0.5)^{5-5}\right] = 31.25 = 31$

$\therefore$ The expected frequencies are 31, 156, 313, 313, 156, 31

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O – E)^2$	$\frac{(O-E)^2}{E}$
40	31	-9	81	2.61
300	156	144	20726	132.86
250	313	-63	3969	12.68

200	313	-113	12769	40.80
130	156	-26	676	4.33
80	31	49	2401	77.45
				χ^2 =270.72

Number of degrees of freedom $\nu = n-1 = 6-1 = 5$

For $\nu = 5$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 11.07$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ Male and female births are not equally probable.

Example 50: Five unbiased dice were thrown 96 times and the number of times 4, 5 or 6 was obtained is given below:

No. of dices showing 4, 5 or 6	0	1	2	3	4	5
Frequency	1	10	24	35	18	8

Fit a suitable distribution and test for the goodness of fit.

Solution : Let us fit a binomial distributions B(X, n, p) to the given data where p be the probability of a male child.

$\therefore P(X = x) = nC_x p^x q^{n-x}, x = 0,1,2,.....n$

p = P(getting 4, 5 or 6 in throw of a die)

= p(4) + p(5) + p(6) = $\frac{1}{6}+\frac{1}{6}+\frac{1}{6} = \frac{3}{6} = \frac{1}{2}$, q = 1- p = $\frac{1}{2}$

H_0 : There is no significant difference between observed frequencies and expected frequencies.

H_1 : There is significant difference between observed frequencies and expected frequencies.

Here n = 5, $\therefore P(X = x) = 5C_x p^x q^{5-x}, x = 0,1,2,.....5$

Total frequency N = 96

The expected frequencies of getting 0, 1, 2, 3, 4, 5 success are given by

N X P(X = 0) = $96\left[5C_0 (0.5)^0 (0.5)^{5-0}\right] = 3$

N X P(X = 1) = $96\left[5C_1 (0.5)^1 (0.5)^{5-1}\right] = 15$

N X P(X = 2) = $96\left[5C_2 (0.5)^2 (0.5)^{5-2}\right] = 30$

N X P(X = 3) = $96\left[5C_3 (0.5)^3 (0.5)^{5-3}\right] = 30$

N X P(X = 4) = $96\left[5C_4 (0.5)^4 (0.5)^{5-4}\right] = 15$

N X P(X = 5) = $96\left[5C_5 (0.5)^5 (0.5)^{5-5}\right] = 3$

$\therefore$ The expected frequencies are 3, 15, 30,30, 15, 3

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$

1 10 ⟩11 35 18 ⟩26 8	3 15 ⟩18 30 15 ⟩18 3	-7 5 8	49 25 64	2.72 0.83 3.55
				χ^2 =7.11

Number of degrees of freedom $\nu = n-1 = 3-1 = 2$

For $\nu = 2$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 5.99$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ There is significant difference between observed frequencies and expected frequencies.

Example 51: Fit a binomial distribution for the following data and also test the goodness of fit.

x	0	1	2	3	4	5	6
f	5	18	28	12	7	6	4

Solution : Let us fit a binomial distribution with parameters n and p to the given data. Here n=6, Mean $\bar{x} = np$

$\therefore P(X = x) = nC_x p^x q^{n-x}, x = 0,1,2,.....n$

x	0	1	2	3	4	5	6	Total
f	5	18	28	12	7	6	4	N=80
fx	0	18	56	36	28	30	24	192

$\bar{x} = \frac{\sum fx}{N} = \frac{192}{80} = 2.4$ and $\bar{x} = np \Rightarrow 2.4 = 6p \Rightarrow p = 0.4$

$\therefore q = 1 - p = 0.6$

Here n = 6, $\therefore P(X = x) = 6C_x (0.4)^x (0.6)^{6-x}, x = 0,1,2,.....6$

The expected frequencies of getting 0, 1, 2, 3, 4, 5, 6 success are given by

N X P(X = 0) = $80\left[6C_0 (0.4)^0 (0.6)^{6-0}\right] = 3.73 = 4$

N X P(X = 1) = $80\left[6C_1 (0.4)^1 (0.6)^{6-1}\right] = 14.93 = 15$

N X P(X = 2) = $80\left[6C_2 (0.4)^2 (0.6)^{6-2}\right] = 24.88 = 25$

N X P(X = 3) = $80\left[6C_3 (0.4)^3 (0.6)^{6-3}\right] = 22.12 = 22$

N X P(X = 4) = $80\left[6C_4 (0.4)^4 (0.6)^{6-4}\right] = 11.06 = 11$

N X P(X = 5) = $80\left[6C_5 (0.4)^5 (0.6)^{6-5}\right] = 2.95 = 3$

N X P(X = 6) = $80\left[6C_6 (0.4)^6 (0.6)^{6-6}\right] = 0.33 = 0$

$\therefore$ The expected frequencies are 4, 15, 25,22, 11, 3, 0

H_0 : There is no significant difference between observed frequencies and expected frequencies.

H_1 : There is significant difference between observed frequencies and expected frequencies.

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
5 18 } 23	4 15 } 19	4	16	0.842
28	25	3	9	0.36
12	22	-10	100	4.545
7 6 } 17 4	11 3 } 14 0	3	9	0.642
				χ^2 =6.39

Number of degrees of freedom $\nu = n - 2 = 4 - 2 = 2$

For $\nu = 2$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 5.99$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

∴ There is significant difference between observed frequencies and expected frequencies.

Example 52: Fit a Poisson distribution for the following data and test the goods of fit.

Values of x	0	1	2	3	4
Frequencies	122	60	15	2	1

Solution : Let us fit a poison distribution with parameters λ to the given data. Here n = 4 and Mean $\bar{x} = \lambda$

Poisson distribution is P(X=x) = $\frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2.....n$

x	0	1	2	3	4	Total
f	122	60	15	2	1	N=200
fx	0	60	30	6	4	100

$\bar{x} = \frac{\sum fx}{N} = \frac{100}{200} = 0.5$ and $\bar{x} = \lambda \Rightarrow \lambda = 0.5$

Poisson distribution is P(X=x) = $\frac{(0.6065)(0.5)^x}{x!}, x = 0,1,2.....4$

$e^{-\lambda} = e^{-0.5} = 0.6065$

N X P(X = 0) = $200\left[\frac{(0.6065)(0.5)^0}{0!}\right] = 121.3 = 121$

N X P(X = 1) = $200\left[\frac{(0.6065)(0.5)^1}{1!}\right] = 60.65 = 61$

N X P(X = 2) = $200\left[\frac{(0.6065)(0.5)^2}{2!}\right] = 15.16 = 15$

N X P(X = 3) = $200\left[\frac{(0.6065)(0.5)^3}{3!}\right] = 2.25 = 3$

N X P(X = 4) = $200\left[\frac{(0.6065)(0.5)^4}{4!}\right] = 0.316 = 0$

$\therefore$ The expected frequencies are 121, 61, 15,3, 0

H_0 : There is no significant difference between observed frequencies and expected frequencies.

H_1 : There is significant difference between observed frequencies and expected frequencies.

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
122	121	1	1	0.0083
60	61	-1	1	0.0164
15, 2, 1 (18)	15, 3, 0 (18)	0	0	0
				χ^2 =0.0247

Number of degrees of freedom $\nu = n - 2 = 3 - 2 = 1$

For $\nu = 1$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 3.84$

Inference : Since the calculated value of χ^2 < the table value of χ^2, H_0 is accepted at 5% level of significance.

$\therefore$ There is no significant difference between observed frequencies and expected frequencies.

Example 53: Fit a poisson distribution for the following data and test the goodness of fit.

X	0	1	2	3	4	5	6
Frequency	56	156	132	92	37	22	5

Solution : Let us fit a poison distribution with parameters λ to the given data. Here n = 6 and Mean $\bar{x} = \lambda$

Poisson distribution is P(X=x) = $\frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2.....n$

x	0	1	2	3	4	5	6	Total
f	56	156	132	92	37	22	5	500
fx	0	156	264	276	148	110	30	984

$\bar{x} = \frac{\sum fx}{N} = \frac{984}{500} = 1.97$ and $\bar{x} = \lambda \Rightarrow \lambda = 1.97$

Poisson distribution is P(X = x) = $\frac{(0.1394)(1.97)^x}{x!}, x = 0,1,2.....6$

$e^{-\lambda} = e^{-1.97} = 0.1395$

N X P(X = 0) = $500\left[\frac{(0.1395)(1.97)^0}{0!}\right] = 69.75 = 70$

N X P(X = 1) = $500\left[\frac{(0.1395)(1.97)^1}{1!}\right] = 137.41 = 137$

N X P(X = 2) = $500\left[\frac{(0.1395)(1.97)^2}{2!}\right] = 135.24 = 135$

N X P(X = 3) = $500\left[\frac{(0.1395)(1.97)^3}{3!}\right] = 88.85 = 89$

N X P(X = 4) = $500\left[\frac{(0.1395)(1.97)^4}{4!}\right] = 43.77 = 44$

N X P(X = 5) = $500\left[\frac{(0.1395)(1.97)^5}{5!}\right] = 17.5 = 18$

N X P(X = 6) = $500\left[\frac{(0.1395)(1.97)^6}{6!}\right] = 5.65 = 6$

$\therefore$ The expected frequencies are 70, 138, 135, 89, 44, 18, 6

H_0 : There is no significant difference between observed frequencies and expected frequencies.

H_1 : There is significant difference between observed frequencies and expected frequencies.

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
56	70	-14	196	2.8
156	137	19	361	2.64
132	135	-3	9	0.067
92	89	3	9	0.1011
37	44	-7	49	1.11
22, 5 ⟩ 27	18, 6 ⟩ 24	3	9	0.375
				$\chi^2 = 7.1$

Number of degrees of freedom $\nu = n - 2 = 6 - 2 = 4$

For $\nu = 4$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 9.48$

Inference : Since the calculated value of χ^2 < the table value of χ^2, H_0 is accepted at 5% level of significance.

$\therefore$ There is no significant difference between observed frequencies and expected frequencies.

4.9 Tests for independence of attributes

For the 2 X 2 contingency table with the cell frequencies a, b, c and d, the χ^2 -value is given by

$$\lambda^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \; where N = a+b+c+d$$

Proof: We take the 2 x 2contingency table

	A	α	Total
B	a	b	B = a + b
β	c	d	β = c + d
Total	A = a + c	α = b + d	N = a + b + c + d

H_0 : The attributes are independent.

H_1 : The attributes are not independent.

$(AB)_0$ = a and $(AB)_e = \frac{A \times B}{N} = \frac{(a+c)(a+b)}{N}$

$(A\beta)_0$ = a and $(A\beta)_e = \frac{A \times \beta}{N} = \frac{(a+c)(c+d)}{N}$

$(\alpha B)_0$ = a and $(\alpha B)_e = \frac{\alpha \times B}{N} = \frac{(b+d)(a+b)}{N}$

$(\alpha\beta)_0$ = a and $(\alpha\beta)_e = \frac{\alpha \times \beta}{N} = \frac{(b+d)(c+d)}{N}$

$$(AB)_0 - (AB)_c = a - \frac{(a+c)(a+b)}{N} = \frac{aN-(a+c)(a+b)}{N}$$

$$= \frac{a(a+b+c+d)-(a^2+ab+ac+bc)}{N} = \frac{ad-bc}{N}$$

$$\frac{\left((AB)_0-(AB)_c\right)^2}{(AB)_c} = \frac{\left(\frac{ad-bc}{N}\right)^2}{\frac{(a+c)(a+b)}{N}} = \frac{(ad-bc)^2}{N(a+c)(a+b)}$$

Similarly, weget $\frac{[(A\beta)_0-(A\beta)_c]^2}{(A\beta)_c} = \frac{(ad-bc)^2}{N(a+c)(c+d)}$

$$\frac{[(\alpha B)_0-(\alpha B)_c]^2}{(\alpha B)_c} = \frac{(ad-bc)^2}{N(b+d)(a+b)}, \frac{[(\alpha\beta)_0-(\alpha\beta)_c]^2}{(\alpha\beta)_c} = \frac{(ad-bc)^2}{N(b+d)(c+d)}$$

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left\{\begin{array}{l}\frac{1}{(a+c)(a+b)}+\frac{1}{(a+c)(c+d)}+\\ \frac{1}{(b+d)(a+b)}+\frac{1}{(b+d)(c+d)}\end{array}\right\}$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left\{\begin{array}{l}\frac{1}{(a+c)}\left(\frac{1}{(a+b)}+\frac{1}{(c+d)}\right)+\\ \frac{1}{(b+d)}\left(\frac{1}{(a+b)}+\frac{1}{(c+d)}\right)\end{array}\right\}$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left\{\begin{array}{l}\frac{1}{(a+c)}\left(\frac{a+b+c+d}{(a+b)(c+d)}\right)+\\ \frac{1}{(b+d)}\left(\frac{a+b+c+d}{(a+b)(c+d)}\right)\end{array}\right\}$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left(\frac{a+b+c+d}{(a+b)(c+d)}\right)\left\{\frac{1}{(a+c)}+\frac{1}{(b+d)}\right\}$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left(\frac{a+b+c+d}{(a+b)(c+d)}\right)\left(\frac{a+b+c+d}{(a+c)(b+d)}\right)$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left(\frac{(a+b+c+d)^2}{(a+b)(a+c)(b+d)(c+d)}\right)$$

$$\chi^2 = \frac{(ad-bc)^2}{N}\left(\frac{(N)^2}{(a+b)(a+c)(b+d)(c+d)}\right)$$

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Example 54: In an experiment on immunization of cattle from tuberculosis the following results were obtained.

	Affected	Not Affected
Inoculated	12	26
Not-inoculated	16	6

Calculate Chi-square and discuss the effect to vaccine in controlling susceptibility to tuberculosis.

Method – I

Solution : Let A and B denote the attribute tuberculosis and vaccine.

A_1- the attribute affected B_1- the attribute inoculated

A_2- the attribute not affected B_2- the attribute not inoculated

H_0: A and B are independent.

H_1: A and B are not independent.

Observed frequency table

B / A	Affected A_1	Not Affected A_2	Total
Inoculated B_1	a 12	b 26	a+ b = 38
Not Inoculated B_2	c 16	d 6	c +d = 22
Total	a + c = 28	b + d = 32	N = 60

The test statistic is

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{60(12\times 6 - 26\times 16)^2}{38\times 28\times 32\times 22} = 9.48$$

$$\gamma = (m-1)(n-1) = (2-1)(2-1) = 1$$

Number of degrees of freedom

For $\nu = 1$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 3.84$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ A and B are not independent.

Method II

H_0: A and B are independent.

H_1: A and B are not independent

We find the expected frequencies. Here it is enough to find one expected frequency, the remaining can be obtained by subtracting from marginal row and column totals.

$$(A_1B_1)_e = \frac{(A_1)(B_1)}{N} = \frac{38\times 28}{60} = 17.73 = 18$$

The expected frequency table

B / A	Affected A_1	Not Affected A_2	Total
Inoculated B_1	a 18	b = 38 – 18=20	a+ b = 38
Not Inoculated B_2	c=28 - 18=10	d =32- 20=12	c +d = 22
Total	a + c = 28	b + d = 32	N = 60

Under H_0, the test statistic is $\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
12	18	- 6	36	2
16	10	6	36	3.6
26	20	6	36	1.8
6	12	- 6	36	3
				$\chi^2 = 10.4$

Number of degrees of freedom $\nu = (m-1)(n-1) = (2-1)(2-1) = 1$

For $\nu = 1$ d.f, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 3.84$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ A and B are not independent.

Example 55: From the following table, test the independent of literacy and smoking habit

	Smokes	Non-smokes	Total
Literates	83	57	140
Illiterates	45	68	113
Total	128	125	253

Solution :

A_1- the attribute smokers B_1- the attribute literacy

A_2- the attribute non smokers B_2- the attribute illiterates.

H_0: Literacy and smoking are independent.

H_1: Literacy and smoking are not independent.

Observed frequency table

B / A	Smokes A_1	Non-Smokes A_2	Total
Literates B_1	a 83	b 57	a+ b = 140
Illiterates B_2	c 45	d 68	c +d = 113
Total	a + c = 128	b + d = 125	N = 253

The test statistic is

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{253(83\times 68 - 45\times 57)^2}{140\times 128\times 113\times 125} = 9.48$$

$$\gamma = (m-1)(n-1) = (2-1)(2-1) = 1$$

Number of degrees of freedom

For $\nu = 1$ degrees of freedom, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 3.84$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

$\therefore$ Literacy and smoking habits are not independent.

Example 56: A sample of 200 persons with a particular disease was selected. Out of these, 100 were given a drug and the others were not given any drug. The results are as follows:

No. of Persons	Drug	No Drug	Total
Cured	65	55	120
Not Cured	35	45	80
Total	100	100	200

Test whether the drug is effective or not.

Solution :

A_1- the attribute Drug B_1- the attribute Cured

A_2- the attribute no Drug B_2- the attribute not Cured

H_0: The drug is not effective.

H_1: The dug is effective.

Observed frequency table

B / A	Smokes A_1	Non-Smokes A_2	Total
Literates B_1	a 65	b 55	a+ b = 120
Illiterates B_2	c 35	d 45	c +d = 80
Total	a + c = 100	b + d = 100	N = 200

The test statistic is

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \qquad = \frac{200(65\times 45 - 55\times 35)^2}{120\times 80\times 100\times 100} = 2.08$$

$$\gamma = (m-1)(n-1) = (2-1)(2-1) = 1$$

Number of degrees of freedom

For $\nu = 1$ d.f, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 3.84$

Inference : Since the calculated value of χ^2 < the table value of χ^2, H_0 is accepted at 5% level of significance.

$\therefore$ The drug is not effective.

Example 57 : Suppose we wish to analyses a set of data to determine if success in medical school is related to ability in Mathematics. Suppose that 150 randomly selected medical students were rated according to their success in medical school and their ability in Mathematics with respect to each of these characteristics the students were rated as low, average or high. The number of students in each category is given in the following table:

	Ability in Mathematics			
Success in Medical School		Low	Average	High
	Low	14	8	5
	Average	12	51	11
	High	7	24	18

On the basis of contingency table, should we conclude that success in medical school is related to ability in Mathematics? Test at the 5% level of significance?

Solution : Let A denote the attribute ability in Mathematics and B denote success in Medical school. We want to test whether A and B are related or not.

H_0: A and B are independent.

H_1: A and B are not independent.

Observed frequency table

B / A	Low A_1	Average A_2	High A_3	Total
Low B_1	14	8	5	27
Average B_2	12	51	11	74
High B_3	7	24	18	49
Total	33	83	34	N = 150

Now, we find only 4 expected frequency and rest can be found from the table

$$(A_1B_1)_e = \frac{(A_1)(B_1)}{N} = \frac{33\times 27}{150} = 5.94 = 6$$

$$(A_1B_2)_e = \frac{(A_1)(B_2)}{N} = \frac{33\times 74}{150} = 16.28 = 16$$

$$(A_2B_1)_e = \frac{(A_2)(B_1)}{N} = \frac{83\times 27}{150} = 14.94 = 15$$

$$(A_2B_2)_e = \frac{(A_2)(B_2)}{N} = \frac{83 \times 74}{150} = 40.9 = 41$$

Expected frequency table

	A_1	A_2	A_3	Total
B_1	6	15	27 – 21 = 6	27
B_2	16	41	74 – 57 = 17	74
B_3	33 – 22 = 11	83 – 56 = 27	49 – 38 = 11	49
Total	33	83	34	150

O	E	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
14	6	8	64	10.7
12	16	-4	16	1
7	11	-4	16	1.45
8	15	-7	49	3.267
51	41	10	100	2.439
24	27	-3	9	0.333
5	6	-1	1	0.167
11	17	-6	36	2.118
18	11	7	49	4.455
				χ^2 = 25.93

Number of degrees of freedom Number of degrees of freedom

$\gamma = (m-1)(n-1) = (3-1)(3-1) = 4$

For $\nu = 4$ d.f, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 9.488$

Inference : Since the calculated value of χ^2 > the table value of χ^2, H_0 is rejected at 5% level of significance.

∴ A and B are not independent.

Example 58: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling into different intelligence level. The results are as follows.

Researchers	Students				
	Below average	Average	Above average	Excellent	Totals
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two teachers are significantly different?

Solution : Let A denote the attribute intelligence level and B be the attribute Research techniques.

H_0: There is no significant difference between the sampling techniques adopted by the two researchers.

H_1: There is significant difference between the sampling techniques adopted by the two researchers.

Observed frequency table

B / A	A_1	A_2	A_3	A_4	Total
B_1	86	60	44	10	200
B_2	40	33	25	2	100
Total	126	93	69	12	N = 300

Now, we find only 4 expected frequency and rest can be found from the table

$$(A_1B_1)_e = \frac{(A_1)(B_1)}{N} = \frac{126 \times 200}{300} = 84$$

$$(A_2B_1)_e = \frac{(A_2)(B_1)}{N} = \frac{93 \times 200}{300} = 62$$

$$(A_3B_1)_e = \frac{(A_3)(B_1)}{N} = \frac{69 \times 200}{300} = 46$$

Expected frequency table

	A_1	A_2	A_3	A_4	Total
B_1	84	62	46	200 – 192 = 8	200
B_2	126 – 84 = 42	93 – 62 = 31	69 – 46 = 23	12 – 8 = 4	100
Total	126	93	69	12	300

O	E	(O – E)	$(O - E)^2$	$\frac{(O-E)^2}{E}$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.5
40	42	-2	4	0.095
33	31	2	4	0.129

25 / 2 \ 27	23 / 4 \ 27	0	0	0
				$\chi^2 =$ 0.924

Number of degrees of freedom Number of degrees of freedom

$\gamma = (m-1)(n-1)-1 = (4-1)(2-1)-1 = 3-1 = 2$

For $\nu = 2$ d.f, the table value of χ^2 at 5% level is $\chi^2_{0.05} = 5.999$

Inference : Since the calculated value of χ^2 < the table value of χ^2, H_0 is accepted at 5% level of significance.

$\therefore$ There is no significant difference between the sampling techniques used by the two researchers.

Summary

- Types of Statistical Hypotheses
- Basic Concepts Concerning Testing Of Hypotheses
- Test for the significant Large Sample tests
- Test for the significant Small Sample tests
- t-Test for paired Observations
- F-Distribution
- Chi-square Distribution for Goodness of fit.
- Testing attributes dependency.

Keywords

- **Null Hypothesis (H0):** A statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.
- **Alternative Hypothesis (H1 or Ha):** A statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters. Alternative hypothesis is created in a negative meaning of the null hypothesis.
- **The level of significance**: It is always some percentage (usually 5%) which should be chosen with great care, thought and reason. In case we take the significance level at 5 per cent, then this implies that H0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H0 is true.
- **Type 1 error**: To reject the null hypothesis when it is true is to make what is known as a type I error. The level at which a result is declared significant is known as the type I error rate, often denoted by α.
- **Type II error:** If we do not reject the null hypothesis when in fact there is a difference between the groups, we make what is known as a type II error. The type II error rate is often denoted as β
- **One- tailed and Two-tailed Tests**: A test of statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one tailed test.

  A test of statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test.
- **Large Sample tests:** When the sample size is n≥30, then apply large sample tests

- **Small Sample tests**: When the sample size is n<30 then apply small sample tests.

SELF-ASSESSMENT QUESTIONS

Short Answer Type Questions

1. The mean breaking strength of the cables supplied by a manufacturer is 1800 with an SD of 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. To test this claim a sample of 50 cables is tested and is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance. (Answer. Z = 3.5)
2. A sample of 900 members has a mean 3.4c.m and S.D. 2.61cm. Is the sample from a large population of mean 3.25cm and S.D. 2.61cm. If the population is normal and the mean is unknown, find the 95% confidence limits for the mean. (Answer. $|Z|$ = 1.7)
3. A company produces two makes of bulbs A and B. 200 bulbs of each make were tested and it was found that the make A had mean life of 2560 hours and S.D 90 hours, whereas make B had 2650 hours mean life and S.D 75 hours. Is there a significant difference between the mean life of two makes? (Answer $|Z|$ = 10.9)
4. A sample of 400 male student of a college is found to have a mean height of 171.38cm. Can it be regarded as a sample from a large population with mean height 171.17 cm and standard deviation 3.3cm. (Answer Z = 1.3)
5. A normal population has standard deviation 30, a sample of 100 has an average 35.5 and another sample of 150 has average 36.2. Is the difference significant? (Answer. $|Z|$ = 1.3)
6. A manufacturer of ball pens claims that a certain pen he manufactures has a mean writing time of 400 pages with a standard deviation of 20 pages. A purchasing agent selects a sample of 100 pens and put them for test. The mean writing life for the sample has 390 pages. Is the difference significant? (Answer. $|Z|$ = 8.8)
7. A sample of 400 students have a mean height 171.38 cms. Can it be reasonably regarded as a sample from a large population with mean height 171.17 cms and standard deviation 3.30cms ?
8. A manufacturer of light bulbs claims that an average 2% of the bulbs manufactured by his firm are defective. A random sample of 400 bulbs contained 13 defective bulbs. On the basis of this sample, can you support the manufacturer's claim at 5% level of Significance? (Answer. $|z|$=1.79)
9. The mean IQ of a sample of 1600 children was 99. Is it a random sample from a population with mean IQ 100 and standard deviation 15.
10. Certain intelligent test was administered for a large group of people and it has been found that the S.D is 36, the test is given to 120 boys with S.D 30. Another group of 125 girls with S.D 33. Does this indicate any significant difference between the two standard deviations?
11. You are given the following information relating to the purchase of bulbs from two manufacturers A and B

Manufacturer	No. of bulbs	Mean life(hrs)	S.D (hrs)
A	50	1980	80
B	70	2010	60

Is the difference between the S.D's significant?

12. A producer confesses that 22% of the items manufactured by him will be defective. To test his claim a random sample of 80 items were selected and 13 items were noted to be defective. Test the validity of the producer's claim at 1% level of significance. (Answer. $|z|$=1.24)
13. A cigarette manufacturing firm claims that its brand A cigarette out sells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another random sample of 100 smokers prefer brand B, test whether the 8% difference is avalid claim. (Answer. $|z|$=1.02)
14. Eight individuals are chosen at random from a population and their heights are found to be in cms 163, 163, 164, 165, 166, 169, 170, 171. In the light of these data discuss the suggestion that the mean height in the universe is 165cm. (Answer. $|t|$= 1.22)
15. The respective heights of six randomly chosen sailors are (in inches) 63,65,68,69,71 and 72. Those of 10 randomly chosen soldiers are 61,62,65,66,69,69,70,71,72 and 73. Discuss in, the light of these data, thrown on the suggestion that sailors are on the average taller than soldiers.

Long Answer type questions

16. In IQ test were administered to 5 persons before and after they were trained. The results are given below.

IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is change in IQ after the training.

17. Two random variables gave the following results.

$n_1 = 10 \qquad \sum (x_i - \bar{x})^2 = 100.4$

$n_2 = 12 \qquad \sum (y_i - \bar{y})^2 = 115.5$

Test whether the difference in variances is significant at 5% level.

18. A set of 5 identical coins is tossed 320 times and the number of heads appearing each time is recorded.

No.of heads	0	1	2	3	4	5
Frequency	14	45	80	112	61	8

Test whether the coins and unbiased at 5% level of significance.

19. In a certain sample of 2000 items, 1400 families are consumers of tea. Out of 1800 Hindu families 1236 families consumes tea. Use chi-square test to test whether there is any

significant difference between consumption of tea among Hindu and Non-Hindu families.

20. Two groups of 100 people each taken for testing the use of a vaccine, 15% contracted the disease out of inoculated persons, While 25 contracted the disease in the other group. Test the efficacy of the vaccine using χ^2 (Answer. $\chi^2 = 3.125$,H_0 accepted)
21. Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in this election. (Answer. $\chi^2 = 10.09$, H_0 rejected)

Votes for area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

22. In a survey of 200 boys of which 75 intelligent, 40 had skilled fathers 85 of the unintelligent boys unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys. (Answer. $\chi^2 = 8.88$, H_0 rejected)
23. The following data test whether there is any relation between sex and preference of colour. (Answer. $\chi^2 = 36.11$, H_0 rejected)

Sex/Colour	Male	Female	Total
Red	10	40	50
White	70	30	100
Green	30	20	50
Total	110	90	20

FURTHER READINGS

1. Devore, J.L, Probability and Statistics for Engineering and Sciences, Cengage Learning, 8th Edition, New Delhi, 2014.
2. Miller and M. Miller, Mathematical Statistics, Pearson Education Inc., Asia 7th Edition, New Delhi, 2011.
3. Richard Johnson, Miller and Freund's Probability and Statistics for Engineer, Prentice Hall of India Private Ltd., 8th Edition, New Delhi, 2011.

UNIT V

Design of Experiments

CONTENTS

Learning Objectives

In this chapter a student has to learn the

- Analysis of Variance
- One-way classification
- Two-way classification (RBD)
- Latin Square Design (LSD) and 2^2 – Factorial Design

Learning Outcomes

- Understand the shortcomings of comparing multiple means as pairs of hypotheses.
- Understand the steps of the ANOVA method and the method's advantages.
- Compare the means of three or more populations using the ANOVA method.
- Calculate pooled standard deviations and confidence intervals as estimates of standard deviations of populations.
- Understand completely randomized and randomized block methods of experimental design and their relation to appropriate ANOVA methods

Overview

In this Unit, you are going to study about Analysis of variance one way, two way, Latin square and 2^2 – Factorial Design**.** Drawing ANOVA tables

5.1 Introduction

Analysis of Variance (ANOVA)

ANOVA was developed in the 1920‟s by R.A. Fisher. Applications: This technique is used when multiple sample cases are there. The significant difference between the means of two samples can be tested through t-test, but the difficulty arises when we are to find the significant difference amongst more than two sample means at the same time

Assumptions of the ANOVA test

Before we can use the one-way ANOVA, we must see if we satisfy some assumptions, just like we had in our previous hypothesis tests:

1. All observations are independent of one another and randomly selected from the population which they represent.
2. The population at each factor level is approximately normal.
3. The variances for each factor level are approximately equal to one another.

The basic principles of design of experiments.

1. Randomization
2. Replication
3. Local control

5.2 One-way classification

A one-way ANOVA has just one independent variable.

For example, difference in IQ can be assessed by Country, and County can have 2, 20, or more different categories to compare.

Working Procedure

Step 1: Find the total number of observations N.

Step 2: Find the total value of all the observations T.

Step 3: Find the correction factor $\frac{T^2}{N}$

Step 4: Calculate $SST = \sum_j \sum_i x_{ij}^2 - \frac{T^2}{N}$

Step 5: Calculate $SSC = \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + ... + \frac{(\sum x_r)^2}{n_r} \right\} - \frac{T^2}{N}$

Where $\sum x_1 =$ total of all values in sample I

$\sum x_2 =$ total of all values in sample II and so on.

Step 6: $SSE = SST - SSC$

Step 7: Find $F = \frac{MSC}{MSE}$ if MSC > MSE or

$F = \frac{MSE}{MSC}$ if MSC > MSE

Step 8 : Find the table value of F for (r – 1, N – r) df or

F for (N – r, r - 1)df at 5% level of significance.

Conclusion: If the Calculate value of F < the table value of F, we accept H_0, otherwise reject H_0.

Example 1 : As part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts. (Coded values) are as follows.

Position 1	90	82	79	98	83	91	
Position 2	105	89	93	104	89	95	86
Position 3	83	89	80	94			

Perform an analysis of variance to test at the 0.05 level of significance whether the differences among the sample means at the three positions are significant.

Solution : We shall reduce the values subtracting 90 from each of the values.

Position 1	Position 2	Position 3			
x_1	x_2	x_3	x_1^2	x_2^2	x_3^2
0	15	-7	0	225	49
-8	-1	-1	64	1	1
-11	3	-10	121	9	100
8	14	4	64	196	16

-7 1	-1 5 -4		49 1	1 25 16	
-17	31	-14	299	473	166

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1 : Not all means same.

N = 6 + 7 + 4 = 17 ; T = -17 + 31 – 14 = 0 ; $\frac{T^2}{N} = 0$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N} = 299 + 473 + 166 - 0 = 938$$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3}\right\} - \frac{T^2}{N} \qquad = \frac{(-17)^2}{6} + \frac{(31)^2}{7} + \frac{(-14)^2}{4} - 0 = 234.45$$

$$SSE = SST - SSC = 938 - 234.45 = 703.55$$

$MSC = \frac{234.45}{2} = 117.23; MSE = \frac{703.55}{14} = 50.25; F = \frac{MSC}{MSE} = 2.33$ ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between samples Within samples (error)	SSC=234.45 SSE=703.55	2 14	MSC=117.23 MSE=50.25	F_c = 2.33
Total	SST=936			

Table value of F(2, 14) at 5% level = 3.74

Inference : Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is no significant difference between the means at the three positions.

Example 2: The following are the numbers of mistakes made in 5 successive days of 4 technicians working for a photographic laboratory:

Technician I	Technician II	Technician III	Technician IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at 0.01level of significance whether the differences among the four sample means can be attributed to chance.

Solution: We shall reduce the values subtracting 10 from each of the values.

x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
-4	4	0	-1	-1	16	16	0	1
4	-1	2	2	7	16	1	4	4
0	2	-3	-2	-3	0	4	9	4
-2	0	5	0	3	4	0	25	0
1	4	1	1	7	1	16	1	1
-1	9	5	0	13	37	37	39	10

H_0 : There is no significant difference between the technicians.

H_1 : There is significant difference between the technicians.

N = 5 + 5 + 5 + 5 = 20 ; T = -1 + 9 + 5 + 0 = 13 ; $\frac{T^2}{N} = \frac{13^2}{20} = 8.45$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N} = 37 + 37 + 39 + 10 - 8.45 = 114.55$$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(-1)^2}{5} + \frac{(9)^2}{5} + \frac{(5)^2}{5} - 0 - 8.45 = 12.95$$

$$SSE = SST - SSC = 114.55 - 12.95 = 101.6$$

Here MSC > MSE, then F = $\frac{MSE}{MSC}$

$$MSC = \frac{12.95}{3} = 4.32; MSE = \frac{101.6}{16} = 5.35; F = \frac{MSE}{MSC} = 5.29$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between samples	SSC=12.95	3	MSC=4.32	
Within samples (error)	SSE=101.6	16	MSE=6.35	F_c = 1.47
Total	SST=114.55			

Table value of F(3, 16) at 1% level = 5.29

Inference : Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is no significant difference between the technicians.

Example 3: A random sample is selected from each of three makes of ropes and their breaking strength (in pounds) are measured with the following results:

I	70	72	75	80	83		
II	100	110	108	112	113	120	107
III	60	65	57	84	87	73	

Test whether the breaking strength of the ropes differs significantly.

Solution : We shall reduce the values subtracting 80 from each of the values.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2
-10	20	-20	100	400	400
-8	30	-25	64	900	225
-5	28	-15	25	784	529
0	32	-23	0	1024	16
3	33	-4	9	1089	49
	40	-7		1600	49
	27	7		729	
		-7			
-20	210	-54	198	6526	1268

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1 : Not all means same.

N = 5 + 7 + 6 = 18 ; T = -20 + 210 -54= 136 ; $\frac{T^2}{N} = \frac{136^2}{18} = 1027.56$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N} = 198 + 6526 + 1268 - 1027.56 = 6964.44$$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3}\right\} - \frac{T^2}{N}$$

$$= \frac{(-20)^2}{5} + \frac{(210)^2}{7} + \frac{(-54)^2}{6} - 1027.56 = 5838.44$$

$$SSE = SST - SSC = 6964.44 - 5838.44 = 1126$$

$$MSC = \frac{SSC}{r-1} = \frac{5838.44}{2} = 2919.22; MSE = \frac{SSE}{N-r} = \frac{1126}{15} = 75.07;\ F = \frac{MSC}{MSE} = 38.89$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio

Between samples Within samples (error)	SSC=5838.44 SSE=1126	2 15	MSC=2919.22 MSE=75.07	F_c = 38.89
Total	SST=6964.44			

Table value of F(2, 15) at 5% level = 3.68

Inference : Since the calculated value of F > the table value of F, H_0 is rejected at 5% level of significance.

$\therefore$ There is significant difference between the means at the three positions.

Example 4: There are three main brands of a certain powder. A set of 110 sample values is examined and found to be allocated among four groups (A, B, C and D) and three brands (I, II, III) as shown under:

Brands	Groups			
	A	B	C	D
I	0	4	8	1
II	5	8	1	5
III	8	1	3	6
		9	1	1
			1	3

Is there any significant difference in brands preference? Answer at 5% level.

Solution: We want to test the difference in brands preference only. So we shall use one-way classification.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2
0	5	8	0	25	64
4	8	19	16	64	361
8	13	11	64	169	121
15	6	13	25	36	169
27	32	51	105	294	715

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1 : Not all means same.

$$N = 4 + 4 + 4 + 4 = 16; T = 27 + 32 + 51 = 110 \quad \frac{T^2}{N} = \frac{110^2}{12} = 1008.33$$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N} = 105 + 294 + 715 - 1008.33 = 105.67$$

$$SSC = \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} \right\} - \frac{T^2}{N}$$

$$= \frac{(27)^2}{4} + \frac{(32)^2}{4} + \frac{(51)^2}{4} - 1008.33 = 80.17$$

$$SSE = SST - SSC = 105.67 - 80.17 = 25.5$$

$$MSC = \frac{SSC}{r-1} = \frac{80.17}{2} = 40.09; MSE = \frac{SSE}{N-r} = \frac{25.5}{9} = 2.83;\ F = \frac{MSC}{MSE} = 14.15$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between samples Within samples (error)	SSC=80.17 SSE=25.5	2 9	MSC=40.09 MSE=2.83	F_c = 14.15
Total	SST=105.67			

Table value of F(2, 9) at 5% level = 4.26

Inference : Since the calculated value of F > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant difference between the means at the three positions.

Example 5: A completely randomized design experiment with 10 plots and 3 treatments gave the following results:

Plot No.	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Analysis the results for treatment effects.

Solution:

A	5	7	3	1
B	4	4	7	
C	3	5	1	

H_0 : There is no significant difference

H_1 : There is significant difference.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2
5	4	3	25	16	9
7	4	5	49	16	25
3	7	1	9	49	1
1			1		
16	15	9	84	81	35

N = 4 + 3 + 3 = 10; T = 16 + 15 + 9 = 40 ; $\frac{T^2}{N} = \frac{40^2}{10} = 160$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N} = 84 + 81 + 35 - 160 = 40$$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3}\right\} - \frac{T^2}{N} \qquad = \frac{(16)^2}{4} + \frac{(15)^2}{3} + \frac{(9)^2}{3} - 160 = 6$$

$$SSE = SST - SSC = 40 - 6 = 34$$

Here MSC > MSE, then F = $\frac{MSE}{MSC}$

$$MSC = \frac{6}{2} = 3; MSE = \frac{34}{7} = 4.86; F = \frac{MSE}{MSC} = 1.62$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between samples	SSC = 6	2	MSC = 3	
Within samples (error)	SSE=34	7	MSE = 4.86	F_c = 1.62
Total	SST=40			

Table value of F(7,2) at 5% level = 19.35

Inference : Since the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

5.3 Two-way classification (RBD)

A two-way ANOVA (are also called factorial ANOVA) refers to an ANOVA using two independent variables. Expanding the example above, a 2-way ANOVA can examine differences in IQ scores (the dependent variable) by Country (independent variable 1) and Gender (independent variable 2). Two-way ANOVA can be used to examine the interaction between the two independent variables. Interactions indicate that differences are not uniform across all categories of the independent variables. For example, females may have higher IQ scores overall compared to males, but this difference could be greater (or less) in European countries compared to North American countries.

Comparisons of Two-way to One-Factor-at-a-Time

1. usually have a smaller total sample size, since you're studying two things at once
2. removes some of the random variability (some of the random variability is now explained by the second factor, so you can more easily find significant difference
3. we can look at interactions between factors (a significant interaction means the effect of one variable changes depending on the level of the other factor).

Working Procedure

Step 1: Find the total number of observations N.

Step 2: Find the total value of all the observations T.

Step 3: Find the correction factor $\frac{T^2}{N}$

Step 4: Calculate $SST = \sum_j \sum_i x_{ij}^2 - \frac{T^2}{N}$

Step 5: Calculate $SSC = \left\{ \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + ... + \frac{(\sum x_r)^2}{n_r} \right\} - \frac{T^2}{N}$

Where $\sum x_1$ = total of all values in sample I

$\sum x_2$ = total of all values in sample II and so on.

$$SSR = \left\{ \frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4^2)}{n_4} \right\} - \frac{T^2}{N}$$

Step 6: $SSE = SST - SSC$ -SSR

Step 7:

$$MSC = \frac{SSC}{c-1} : MSR = \frac{SSR}{r-1} \quad MSE = \frac{SSE}{(c-1)(r-1)}$$

Step 8: Find

$$F_R = \frac{MSR}{MSE} \quad F_C = \frac{MSC}{MSE}$$

Step 8 : Find the table value of F for (c – 1,r – 1) df or

F for (c – 1, r - 1) df at 5% level of significance.

Conclusion: If the Calculate value of F < the table value of F, we accept H_0, otherwise reject H_0.

Example 6 : An experiment was designed to study the performance of 4different detergents for cleaning fuel injectors. The following cleanness readings were obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines:

	Engine 1	Engine 2	Engine 3	Total
Detergent A	45	43	51	139
Detergent B	47	46	52	145
Detergent C	48	50	55	153
Detergent D	42	37	49	128
Total	182	176	207	565

Perform the ANOVA and test at 0.01level of significance whether there are differences in the detergents or in the engines.

Solution : We shall reduce the values subtracting 50 from each of the values.

In this problem the data is given according to two factors, detergent and engine. So, we do 2-way analysis of variance.

Engine|Detergent	x_1	x_2	x_3	Total	x_1^2	x_2^2	x_3^2
y_1	-5	-7	1	-11	25	49	1
y_2	-3	-4	2	-5	9	16	4
y_3	-2	0	5	3	4	0	25
y_4	-8	-13	-1	-22	64	169	1
Total	-18	-24	7	-35	102	234	31

H_0 : There is no difference between the engines and between the detergents.

H_1 : There is difference between the engines and between the detergents.

N = 4 + 4 + 4 = 12 ; T = - 35 ; $\frac{T^2}{N} = \frac{(-35)^2}{12} = 102.08$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N}$

$= 102 + 234 + 31 - 102.08 = 264.92$ $SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3}\right\} - \frac{T^2}{N}$

$= \frac{(-18)^2}{4} + \frac{(-24)^2}{4} + \frac{(7)^2}{4} - 102.08 = 135.17$

$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4^2)}{n_4}\right\} - \frac{T^2}{N}$

$= \frac{(-11)^2}{3} + \frac{(-5)^2}{3} + \frac{(3)^2}{3} + \frac{(-22)^2}{3} - 102.08 = 110.92$

$SSE = SST - SSC - SSR = 264.92 - 135.17 - 110.92 = 18.83$

$MSC = \frac{SSC}{c-1} = \frac{135.17}{2} = 67.59; MSR = \frac{SSR}{r-1} = \frac{110.92}{3} = 36.97;$

$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{18.83}{6} = 3.14;$

$F_C = \frac{MSC}{MSE} = 21.52, F_R = \frac{MSR}{MSE} = 11.77$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=135.17	2	MSC=67.59	F_C = 21.52
Between rows	SSR=110.92	3	MSR=36.97	F_R = 11.77
Residual (error)	SSE=18.83	6	MSE=3.14	
Total	SST=264.92			

Table value of F_C (2,6) at 1% level = 10.92

Table value of F_R (3,6) at 1% level = 9.78

Inference: Since the calculated value of F > the table value of F, H_0 is rejected at 1% level of significance.

∴ There is significant difference between the engines and between the detergents.

Example 7: Four different, though supposedly equivalent, forms of standardized reading achievement test were given to each of 5 students, and the following are the scores which they obtained.

	Student 1	Student 2	Student 3	Student 4	Student 5
Form A	75	73	59	69	84
Form B	83	72	56	70	92
Form C	86	61	53	72	88
Form D	73	67	62	79	95

Perform a two-way analysis of variance to test at the level of significance α = 0.01 whether it is reasonable to treat the 4 forms as equivalent.

Solution: We shall reduce the values subtracting 70 from each of the values.

In this problem the data is given according to two factors, Student's and Forms. So, we do 2-way analysis of variance.

	Student										
Form	x_1	x_2	x_3	x_4	x_5	Total	x_1^2	x_2^2	x_3^2	x_4^2	x_5^2
y_1	5	3	-11	-1	14	10	25	9	121	1	196
y_2	13	2	-14	0	22	23	169	4	196	0	484
y_3	16	-9	-17	2	18	10	256	81	289	4	324
y_4	3	-3	-8	9	25	26	9	9	64	81	625
Total	37	-7	-50	10	79	69	459	103	670	86	1629

H_0 : There is no difference between the student's and between the forms.

H_1 : There is difference between the student's and between the forms.

N = 4 + 4 + 4 + 4 + 4 = 20 ; T = 69 ; $\frac{T^2}{N} = \frac{(69)^2}{20} = 238.05$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2 + \sum x_5^2\right] - \frac{T^2}{N}$$

$$= 459 + 103 + 670 + 86 + 1629 - 238.05 = 2708.85$$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} + \frac{(\sum x_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(37)^2}{4} + \frac{(-7)^2}{4} + \frac{(-50)^2}{4} + \frac{(10)^2}{4} + \frac{(79)^2}{4} - 238.05 = 2326.7$$

$$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4^2)}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(10)^2}{5} + \frac{(23)^2}{5} + \frac{(10)^2}{5} + \frac{(26)^2}{5} - 238.05 = 42.95$$

$$SSE = SST - SSC - SSR = 2708.85 - 2326.7 - 42.95 = 339.2$$

$$MSC = \frac{SSC}{c-1} = \frac{2326.7}{4} = 581.68; MSR = \frac{SSR}{r-1} = \frac{42.95}{3} = 14.32;$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{339.2}{12} = 28.26;$$

$$F_C = \frac{MSC}{MSE} = 20.58, F_R = \frac{MSR}{MSE} = 1.97$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns Between rows Residual (error)	SSC=2326.7 SSR=42.95 SSE=339.2	4 3 12	MSC=581.68 MSR=14.32 MSE=28.26	F_C = 20.58 F_R = 1.97
Total	SST=2708.85			

Table value of F_C (4,12) at 1% level = 5.41

Table value of F_R (12,3) at 1% level = 27.05

Inference : Since the calculated value of F_C > the table value of F, H_0 is rejected at 1% level of significance.

Since the calculated value of F_R < the table value of F, H_0 is accepted at 1% level of significance.

∴ There is no significant difference between the forms.

Example 8: A company appoints 4 sales men A, B, C and D and observes their sales in 3 seasons. Summer, Winter and Monsoon. The figures (in lakes of Rs.) are given in the following table.

Sales men

Season		A	B	C	D
	Summer	45	40	38	37
	Winter	43	41	45	38
	Monsoon	39	39	41	41

Carry out an analysis of variance.

Solution :

We shall reduce the values subtracting 40 from each of the values.

In this problem the data is given according to two factors, season and salesmen. So, we do 2-way analysis of variance.

Salesmen

Seasons	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	5	0	-2	-3	0	25	0	4	9
y_2	3	1	5	-2	7	59	1	25	4
y_3	-1	-1	1	1	0	91	1	51	1
Total	7	0	4	-4	7	35	2	30	14

H_0 : There is no significant difference between the salesmen and between seasons .

H_1 : There is significant difference between the salesmen and between seasons .

N = 3 + 3 + 3+3 = 12 ; T = 7 ; $\frac{T^2}{N} = \frac{(7)^2}{12} = 4.08$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 35 + 2 + 30 + 14 - 4.08 = 76.92$ $SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$

$= \frac{(49)^2}{3} + \frac{(0)^2}{3} + \frac{(4)^2}{3} + \frac{(-4)^2}{3} - 4.08 = 22.92$

$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3}\right\} - \frac{T^2}{N}$

$= \frac{(0)^2}{4} + \frac{(7)^2}{4} + \frac{(0)^2}{4} - 4.08 = 8.17$

$$SSE = SST - SSC - SSR = 76.92 - 22.92 - 8.17 = 45.83$$

$$MSC = \frac{SSC}{c-1} = \frac{22.92}{3} = 7.64; MSR = \frac{SSR}{r-1} = \frac{8.17}{2} = 4.08; \; MSE = \frac{SSE}{(c-1)(r-1)} = \frac{45.83}{6} = 7.64$$

$$F_C = \frac{MSC}{MSE} = 1, F_R = \frac{MSE}{MSR} = 1.87,$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=22.92	3	MSC=7.64	F_C = 1
Between rows	SSR=8.17	2	MSR=4.08	F_R = 1.87
Residual (error)	SSE=45.83	6	MSE=7.64	
Total	SST=76.92			

Table value of F_C (3,6) at 5% level = 4.76

Table value of F_R (6,2) at 5% level = 19.33

Inference : In both cases the calculated value of F < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is no significant difference between the salesmen and between seasons.

Example 9: The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.

Machine Type

Workers	A	B	C	D
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Test whether the five men differ with respect to mean productivity and test whether the mean productivity is the same for the four different machine types.

Solution : We shall reduce the values subtracting 40 from each of the values.

In this problem the data is given according to two factors, workers and machine types. So, we do 2-way analysis of variance.

Machine Type

workers	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	4	-2	7	-4	5	16	4	49	16
y_2	6			3	21	36	0		9

y_3 y_4 y_5	-6 3 -2	0 -4 -2 2	1 2 4 6 9	-8 -7 -1	-14 0 8	36 9 4	16 4 4	14 4 16 36 81	64 49 1
Total	5	-6	38	-17	20	101	28	326	139

H_0 : There is no significant difference between the workers and between machine types.

H_1 : There is significant difference between the workers and between machine types.

N = 20 ; T = 20 ; $\frac{T^2}{N} = \frac{(20)^2}{20} = 20$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 101 + 28 + 326 + 139 - 20 = 574$ $SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$

$= \frac{(25)^2}{5} + \frac{(36)^2}{5} + \frac{(38)^2}{5} + \frac{(-17)^2}{5} - 20 = 338.8$

$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4} + \frac{(\sum y_5)^2}{n_5}\right\} - \frac{T^2}{N}$

$= \frac{(5)^2}{4} + \frac{(21)^2}{4} + \frac{(-14)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4} - 20 = 161.5$

$SSE = SST - SSC - SSR = 574 - 161.5 - 338.8 = 73.7$

$MSC = \frac{SSC}{c-1} = \frac{338.8}{3} = 112.93; MSR = \frac{SSR}{r-1} = \frac{161.5}{4} = 40.38;$

$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{73.7}{12} = 6.14$

$F_C = \frac{MSC}{MSE} = 18.39, F_R = \frac{MSR}{MSE} = 6.57$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=338.8	3	MSC=112.93	F_C = 18.39
Between rows	SSR=161.5	4	MSR=40.38	F_R = 6.57
Residual (error)	SSE=83.7	12	MSE=6.14	
Total	SST=584			

Table value of F_C (3,4) at 5% level = 3.49

Table value of F_R (4,12) at 5% level = 3.26

Inference : In both cases the calculated value of F > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant difference between the workers and between machine types.

Example 10: A laboratory technician measures the breaking strength of each of 5 kinds of linen threads by using 4different measuring instruments, and obtains the following results, in ounces.

	I_1	I_2	I_3	I_4
Thread 1	20.9	20.4	19.9	21.9
Thread 2	2.5	26.2	27	24.8
Thread 3	25.5	23.1	21.5	24.4
Thread 4	24.8	21.2	23.5	25.7
Thread 5	19.6	21.2	22.1	22.1

Solution : We shall reduce the values subtracting 22.1 from each of the values.

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	-1.2	-1.7	-2.2	-0.2	-5.3	1.44	2.89	4.84	0.04
y_2	2.9	4.1	4.9	2.7	14.6	8.41	16.81	24.01	7.29
y_3	3.4	1	-0.6	2.3	6.1	11.56	1	0.36	5.29
y_4	2.7	-0.9	1.4	3.6	6.8	7.29	0.81	1.96	12.96
y_5	-2.5	-0.9	0	0	-3.4	6.25	0.81	0	0
	5.3	1.6	3.5	8.4	18.8	34.95	22.32	31.17	25.58

H_0 : There is no significant in breaking strength of various threads.

H_1 : There is significant in breaking strength of various threads.

N = 20 ; T = 18.8 ; $\frac{T^2}{N} = \frac{(18.8)^2}{20} = 17.67$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 34.95 + 22.32 + 31.17 + 25.58 + -17.67 = 96.35$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(5.3)^2}{5} + \frac{(1.6)^2}{5} + \frac{(3.5)^2}{5} + \frac{(8.4)^2}{5} - 17.67 = 5.02$$

$$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4} + \frac{(\sum y_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(-5.3)^2}{4} + \frac{(14.6)^2}{4} + \frac{(6.1)^2}{4} + \frac{(6.8)^2}{4} + \frac{(-3.4)^2}{4} - 17.67 = 66.4$$

$$SSE = SST - SSC - SSR = 96.35 - 5.02 - 66.4 = 24.93$$

$$MSC = \frac{SSC}{c-1} = \frac{5.02}{3} = 1.67; MSR = \frac{SSR}{r-1} = \frac{66.4}{4} = 16.6;\ MSE = \frac{SSE}{(c-1)(r-1)} = \frac{24.93}{12} = 2.08$$

$$F_C = \frac{MSE}{MSC} = 1.25, MSE \text{ f } MSC; F_R = \frac{MSR}{MSE} = 7.98$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=5.02	3	MSC=1.67	F_C = 1.25
Between rows	SSR=66.4	4	MSR=16.6	F_R = 7.98
Residual (error)	SSE=24.93	12	MSE=2.08	
Total	SST=96.35			

Table value of F_C (3,4) at 5% level = 3.49

Table value of F_R (4,12) at 5% level = 3.26

Inference : In both cases the calculated value of F > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant difference between the workers and between machine types.

Inference : Since the calculated value of F_C < the table value of F, H_0 is accepted at 5% level of significance.

Since the calculated value of F_R > the table value of F, H_0 is rejected at 5% level of significance.

5.4 Three factor classification or Latin Square Design (LSD)

A Latin square design is a method of placing treatments so that they appear in a balanced fashion within a square block or field. Treatments appear once in each row and column. Replicates are also included in this design.

• Treatments are assigned at random within rows and columns, with each treatment once per row and once per column.

• There are equal numbers of rows, columns, and treatments.

• Useful where the experimenter desires to control variation in two different directions The Latin square design, perhaps, represents the most popular alternative design when two (or more) blocking factors need to be controlled for. A Latin square design is actually an extreme example of an incomplete block design, with any combination of levels involving the two blocking factors assigned to one treatment only, rather than to all!

The ANOVA for a Latin Square Design

• Degrees of freedom (df): (Treatment df) = (Row df) = (Column df) = p – 1

• SStrt = the treatment sum of squares

MStrt = the treatment mean square = SStrt/(p − 1)

• SSrow = the sum of squares for rows

MSrow = the mean square for rows = SSrow/(p − 1)

• SScol = the sum of squares for columns

MScol = the mean square for columns = SScol/(p − 1)

• SSE = the error sum of squares The SSE degrees of freedom = (p − 1)(p − 2)

MSE = the mean square error = SSE/[(p − 1)(p − 2)]

• SStotal = the corrected total sum of squares

The SStotal degrees of freedom = N − 1 = p 2 − 1

The total sum of squares for the latin square design is partitioned into 4 components:

SStotal = SSrow + SStrt + SScol + SSE

Formulas to calculate SStotal, SSrow, SStrt and SScol:

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ijk} - \overline{y}_{..})^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} y_{ijk}^2 - \frac{y_{...}^2}{p^2} \qquad SS_{row} = \sum_{i=1}^{p} p(\overline{y}_{i..} - \overline{y}_{..})^2 = \sum_{i=1}^{p} \frac{R_i^2}{p} - \frac{y_{...}^2}{p^2}$$

$$SS_{trt} = \sum_{j=1}^{p} p(\overline{y}_{.j.} - \overline{y}_{..})^2 = \sum_{j=1}^{p} \frac{T_j^2}{p} - \frac{y_{...}^2}{p^2} \qquad SS_{col} = \sum_{k=1}^{p} p(\overline{y}_{..k} - \overline{y}_{..})^2 = \sum_{k=1}^{p} \frac{C_k^2}{p} - \frac{y_{...}^2}{p^2}$$

$$SS_E = SS_{total} - SS_{row} - SS_{trt} - SS_{col} \qquad \frac{y_{...}^2}{p^2} = \frac{y_{...}^2}{N} = \text{the correction factor.}$$

Latin Square Design ANOVA Table

Source of Variation	Sum of Squares	d.f.	Mean Square	F Ratio
Treatments	SS_{trt}	$p-1$	MS_{trt}	$\frac{MS_{trt}}{MS_E}$
Rows	SS_{row}	$p-1$	MS_{row}	
Columns	SS_{col}	$p-1$	MS_{col}	
Error	SS_E	$(p-1)(p-2)$	MS_E	
Total	SS_{total}	p^2-1		

Example 11: A farmer wishes to test the effects of four different fertilizers A, B, C, D on the yield of wheat. In order to eliminate sources of error due to variability in soil fertility, he uses

the fertilizers in a Latin square arrangement as indicated with following table, where the numbers indicate yields in bushels per unit area.

A 18	C 21	D 25	B 11
D 22	B 12	A 15	C 19
B 15	A 20	C 23	D 24
C 22	D 21	B 10	A 17

Perform an analysis of variance to determine if there is a significant difference between the fertilizers at $\alpha = 0.05$ levels of significance.

Solution : We shall reduce the values subtracting 15 from each of the values

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	A 3	C 6	D 10	B -4	15	9	36	100	16
y_2	D 7	B -3	A 0	C 4	8	49	9	0	16
y_3	B 0	A 5	C 8	D 9	22	0	25	64	81
y_4	C 7	D 6	B -5	A 2	10	49	36	25	4
	17	14	13	11	55	107	106	189	117

H_0 : There is no difference between rows between columns and between treatments.

H_1 : Not all equal.

N = 16 ; T = 55 ; $\frac{T^2}{N} = \frac{(55)^2}{16} = 189.06$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 107 + 106 + 189 + 117 - 189.06 = 329.94$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(17)^2}{4} + \frac{(14)^2}{4} + \frac{(13)^2}{4} + \frac{(11)^2}{4} - 189.06 = 4.69$$

$$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(15)^2}{4} + \frac{(8)^2}{4} + \frac{(22)^2}{4} + \frac{(10)^2}{4} - 189.06 = 29.19$$

To find SSK, we arrange the data according to the letters row wise

A	B	C	D
3	-4	6	10
0	-3	4	7
5	0	8	9
2	-5	7	6
10	-12	25	32

$$SSK = \frac{(10)^2}{4} + \frac{(-12)^2}{4} + \frac{(25)^2}{4} + \frac{(32)^2}{4} - 189.06 = 284.19$$

$$SSE = SST - SSC - SSR - SSK = 329.94 - 4.69 - 29.19 - 284.19 = 11.87$$

$$MSC = \frac{SSC}{n-1} = \frac{4.69}{3} = 1.56; MSR = \frac{SSR}{n-1} = \frac{29.19}{3} = 9.73;$$

$$MSK = \frac{SSK}{n-1} = \frac{284.19}{3} = 94.73; MSE = \frac{SSE}{(n-1)(n-2)} = \frac{11.87}{6} = 1.98$$

$$F_C = \frac{MSE}{MSC} = \frac{1.98}{1.56} = 1.27, MSE \text{ f } MSC; F_R = \frac{MSR}{MSE} = \frac{9.73}{1.978} = 4.92$$

$$F_T = \frac{MSK}{MSE} = \frac{94.73}{1.98} = 47.89$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=4.69	3	MSC=1.56	F_C = 1.27
Between rows	SSR=29.19	3	MSR=9.73	F_R = 4.92
Between Treatments	SSK=284.19	3	MSK=94.73	F_T = 47.89
Residual (error)	SSE=11.87	6	MSE=1.98	
Total	SST=329.94			

Table value of F_C (6,3) at 5% level = 8.94

Table value of F_R (3,6) at 5% level = 4.76

Table value of F_T (3,6) at 5% level = 4.76

Inference: calculated value of F_C < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is significant difference between columns so far as fertility is concerned.

calculated value of F_R > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant in fertility from row to row.

calculated value of F_T > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant difference between the fertilizers.

Example 12: Analyses the variance in the latin square of yields in (kgs) of paddy where P, Q, R, S denote the different methods of cultivation.

S 122	P 121	R 123	Q 122

Q 124	R 123	P 122	S 125
P 120	Q 119	S 120	R 121
R 122	S 123	Q 121	P 122

Examine whether the different methods of cultivation have given significantly different yields.

Solution :

We shall reduce the values subtracting 120 from each of the values.

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	A 2	P 1	R 3	Q 2	8	4	1	9	4
y_2	Q 4	R 3	P 2	S 5	14	16	9	4	25
y_3	P 0	Q -1	S 0	R 1	0	0	1	0	1
y_4	R 2	S 3	Q 1	P 2	8	4	9	1	4
	8	6	6	10	30	24	20	14	34

H_0 : There is no difference between rows between columns and between the methods of cultivation.

H_1 : Not all equal.

N = 16 ; T = 30 ; $\frac{T^2}{N} = \frac{(30)^2}{16} = 56.25$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 24 + 20 + 14 + 34 - 56.25 = 35.75$ $SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$

$= \frac{(8)^2}{4} + \frac{(6)^2}{4} + \frac{(6)^2}{4} + \frac{(10)^2}{4} - 56.25 = 2.75$

$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4}\right\} - \frac{T^2}{N}$

$= \frac{(8)^2}{4} + \frac{(14)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4} - 56.25 = 24.75$

To find SSK, we arrange the data according to the letters row wise

P	Q	R	S
1	2	3	2
2	4	3	5
0	-1	1	0
2	1	2	3
5	6	9	10

$SSK = \frac{(5)^2}{4} + \frac{(6)^2}{4} + \frac{(9)^2}{4} + \frac{(10)^2}{4} - 56.25 = 4.25$

$$SSE = SST - SSC - SSR - SSK = 35.75 - 2.75 - 24.75 - 4.25 = 4$$

$$MSC = \frac{SSC}{n-1} = \frac{2.75}{3} = 0.92; MSR = \frac{SSR}{n-1} = \frac{24.75}{3} = 8.25;$$

$$MSK = \frac{SSK}{n-1} = \frac{4.25}{3} = 1.42; MSE = \frac{SSE}{(n-1)(n-2)} = \frac{4}{6} = 0.67$$

$$F_C = \frac{MSC}{MSE} = \frac{0.92}{0.67} = 1.37; F_R = \frac{MSR}{MSE} = \frac{8.25}{0.67} = 12.31 \; F_T = \frac{MSK}{MSE} = \frac{1.42}{0.67} = 2.12$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=2.75	3	MSC=0.92	F_C = 1.37
Between rows	SSR=24.75	3	MSR=8.25	F_R = 12.31
Between Treatments	SSK=4.25	3	MSK=1.42	F_T = 2.12
Residual (error)	SSE=4	6	MSE=0.67	
Total	SST=35.75			

Table value of F_C (3,6) at 5% level = 4.76

Table value of F_R (3,6) at 5% level = 4.76

Table value of F_T (3,6) at 5% level = 4.76

Inference : calculated value of F_C < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is significant difference between columns of plots in yield.

calculated value of F_R > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant differences between rows of plots in yield.

calculated value of F_T < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is significant difference between the different methods of cultivation in the yield of paddy.

Example 13: Set up the analysis of variance for the following results of a Latin square design. Use $\alpha = 0.01$ level of significance.

A 12	C 19	B 10	D 8
C 18	B 12	D 6	A 7
B 22	D 10	A 5	C 21
D 12	A 7	C 27	B 17

Solution :

H_0 : There is no significant difference between rows, between columns and between the treatments.

H_1 : There is significant difference between rows, between columns and between the treatments.

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
y_1	A 12	C 19	B 10	D 8	49	144	361	100	64
y_2	C 18	B 12	D 6	A 7	43	324	144	36	49
y_3	B 22	D 10	A 5	C 21	58	484	100	25	441
y_4	D 12	A 7	C 27	B 17	63	144	49	729	289
	64	48	48	53	213	1096	654	890	843

N = 16 ; T = 213 ; $\frac{T^2}{N} = \frac{(213)^2}{16} = 2835.56$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N}$

$= 1096 + 654 + 890 + 843 - 2835.56 = 647.44$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(64)^2}{4} + \frac{(48)^2}{4} + \frac{(48)^2}{4} + \frac{(53)^2}{4} - 2835.56 = 42.69$$

$$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4}\right\} - \frac{T^2}{N}$$

$$= \frac{(49)^2}{4} + \frac{(43)^2}{4} + \frac{(53)^2}{4} + \frac{(63)^2}{4} - 2835.56 = 60.19$$

To find SSK, we arrange the data according to the letters row wise

A	B	C	D
12	10	19	8
7	12	18	6
5	22	21	10
7	17	27	12
31	61	85	36

$$SSK = \frac{(31)^2}{4} + \frac{(61)^2}{4} + \frac{(85)^2}{4} + \frac{(36)^2}{4} - 2835.56 = 465.19$$

$$SSE = SST - SSC - SSR - SSK = 647.44 - 42.69 - 60.19 - 465.19 = 79.37$$

$$MSC = \frac{SSC}{n-1} = \frac{42.69}{3} = 14.23; MSR = \frac{SSR}{n-1} = \frac{60.19}{3} = 20.06;$$

$$MSK = \frac{SSK}{n-1} = \frac{465.19}{3} = 155.06; MSE = \frac{SSE}{(n-1)(n-2)} = \frac{79.37}{6} = 13.23$$

$$F_C = \frac{MSC}{MSE} = \frac{14.23}{13.23} = 1.08; F_R = \frac{MSR}{MSE} = \frac{20.06}{13.23} = 1.52 \; F_T = \frac{MSK}{MSE} = \frac{155.06}{13.23} = 11.72$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=42.69	3	MSC=14.23	F_C = 1.08
Between rows	SSR=60.19	3	MSR=20.06	F_R = 1.52
Between Treatments	SSK=465.19	3	MSK=155.06	F_T = 11.72
Residual (error)	SSE=79.37	6	MSE=13.23	
Total	SST=647.44			

Table value of F_C (3,6) at 1% level = 9.78

Table value of F_R (3,6) at 1% level = 9.78

Table value of F_T (3,6) at 1% level = 9.78

Inference: calculated value of F_C < the table value of F, H_0 is accepted at 1% level of significance.
calculated value of F_R < the table value of F, H_0 is accepted at 1% level of significance.

∴ There is no significant differences between rows and columns.

calculated value of F_T > the table value of F, H_0 rejected at 1% level of significance.

∴ The treatments are significantly different.

Example 14: In a 5X5 Latin square experiment, the data collected is given in the matrix below. Yield per plot is given in quintals for the five different cultivation treatments A, B, C, D and E. Perform the analysis of variance.

A48	E66	D56	C52	B61
D64	B62	A50	E64	C63
B69	A53	C60	D61	E67
C57	D58	E67	B65	A55
E67	C57	B66	A60	D57

Solution : We shall reduce the values subtracting 60 from each of the values.

	x_1	x_2	x_3	x_4	x_5		x_1^2	x_2^2	x_3^2	x_4^2	x_5^2
y_1	A-12	E6	D-4	C-8	B1	-17	144	36	16	64	1
y_2	D4	B2	A-10	E4	C3	3	16	4	100	16	9
y_3	B9	A-7	C0	D1	E7	10	81	49	0	1	7
y_4	C-3	D-2	E7	B5	A-5	2	9	4	49	25	25
y_5	E7	C-3	B6	A0	D-3	7	49	9	36	0	9
	5	-4	-1	2	3	5	299	102	201	106	51

H_0 : There is no difference between rows between columns and between treatments.

H_1 : Not all equal.

N = 25; T = 5 ; $\frac{T^2}{N} = \frac{(5)^2}{25} = 1$ $SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2 + \sum x_5^2\right] - \frac{T^2}{N}$

$= 299 + 102 + 201 + 106 + 51 - 1 = 758$

$$SSC = \left\{\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \frac{(\sum x_4)^2}{n_4} + \frac{(\sum x_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(5)^2}{5} + \frac{(-4)^2}{5} + \frac{(-1)^2}{5} + \frac{(2)^2}{5} + \frac{(3)^2}{5} - 1 = 10$$

$$SSR = \left\{\frac{(\sum y_1)^2}{n_1} + \frac{(\sum y_2)^2}{n_2} + \frac{(\sum y_3)^2}{n_3} + \frac{(\sum y_4)^2}{n_4} + \frac{(\sum y_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(-17)^2}{5} + \frac{(3)^2}{5} + \frac{(10)^2}{5} + \frac{(2)^2}{5} + \frac{(7)^2}{5} - 1 = 89.2$$

To find SSK, we arrange the data according to the letters row wise

A	B	C	D	E
-12	1	-8	-4	6
-10	2	3	4	4
-7	9	0	1	7
-5	5	-3	-2	7
0	6	-3	-3	7
-34	23	-11	-4	31

$SSK = \frac{(-34)^2}{5} + \frac{(23)^2}{5} + \frac{(-11)^2}{5} + \frac{(-4)^2}{5} + \frac{(31)^2}{5} - 1 = 555.6$ $SSE = SST - SSC - SSR - SSK$

$= 758 - 10 - 89.2 - 555.6 = 103.2$ $MSC = \frac{SSC}{n-1} = \frac{10}{4} = 2.5; MSR = \frac{SSR}{n-1} = \frac{89.2}{4} = 22.3;$

$MSK = \frac{SSK}{n-1} = \frac{555.6}{4} = 138.9; MSE = \frac{SSE}{(n-1)(n-2)} = \frac{103.2}{12} = 8.6$

$$F_C = \frac{MSE}{MSC} = \frac{8.6}{2.5} = 3.44, MSE \text{ f } MSC; F_R = \frac{MSR}{MSE} = \frac{22.3}{8.6} = 2.59 \; F_T = \frac{MSK}{MSE} = \frac{138.9}{8.6} = 16.15$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=10	4	MSC=2.5	F_C = 3.44
Between rows	SSR=89.	4	MSR=22.3	F_R = 2.59
Between Treatments	2 SSK=555	4 12	MSK=138.9 MSE=8.6	F_T = 16.15
Residual (error)	.6 SSE=103 .2			
Total	SST=758			

Table value of F_C (12,4) at 5% level = 5.91

Table value of F_R (4,12) at 5% level = 3.26

Table value of F_T (4,12) at 5% level = 3.26

Inference : calculated value of F_C < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is no significant difference between columns.

calculated value of F_R < the table value of F, H_0 is accepted at 5% level of significance.

∴ There is no significant difference between rows.

calculated value of F_T > the table value of F, H_0 is rejected at 5% level of significance.

∴ There is significant difference between the treatments.

Example15: The figures in the following 5X5 Latin square are the numbers of minutes, engines E_1, E_2, E_3, E_4, and E_5, tuned up by mechanics M_1, M_2, M_3, M_4, and M_5, run with a gallon of fuel A, B, C, D and E.

	E_1	E_2	E_3	E_4	E_5
M_1	A31	B24	C20	D20	E18
M_2	B21	C27	D23	E25	A31
M_3	C21	D27	E25	A29	B21
M_4	D21	E25	A33	B25	C22
M_5	E21	A37	B24	C24	D20

Use the level of significance $\alpha = 0.01$to test

i) The H_0 that there is no difference in the performance of the five engines.
ii) H_0 that the persons who tuned up these engines have no effect on their performance.
iii) H_0 that the engines perform equally well with each of the fuels.

Solution : We shall reduce the values subtracting 25 from each of the values.

	E_1	E_2	E_3	E_4	E_5		E_1^2	E_2^2	E_3^2	E_4^2	E_5^2
M_1	A6	B-1	C-5	D-5	E-7	-12	36	1	25	25	49
M_2	B-4	C2	D-2	E0	A6	2	16	4	4	0	36
M_3	C-4	D2	E0	A4	B-4	-2	16	4	0	16	16
M_4	D-4	E0	A8	B0	C-3	1	16	0	64	0	9
M_5	E-4	A12	B-1	C-1	D-5	1	16	144	1	1	25
	-10	15	0	-2	-13	-10	100	153	94	42	135

H_0 : There is no difference between rows between columns and between treatments.

H_1 : Not all equal.

N = 25; T = -10 ; $\frac{T^2}{N} = \frac{(-10)^2}{25} = 4$ $SST = \left[\sum E_1^2 + \sum E_2^2 + \sum E_3^3 + \sum E_4^2 + \sum E_5^2\right] - \frac{T^2}{N}$

$= 100 + 153 + 94 + 42 + 135 - 4 = 520$

$$SSC = \left\{\frac{(\sum E_1)^2}{n_1} + \frac{(\sum E_2)^2}{n_2} + \frac{(\sum E_3)^2}{n_3} + \frac{(\sum E_4)^2}{n_4} + \frac{(\sum E_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(-10)^2}{5} + \frac{(15)^2}{5} + \frac{(0)^2}{5} + \frac{(-2)^2}{5} + \frac{(-13)^2}{5} - 4 = 95.6$$

$$SSR = \left\{\frac{(\sum M_1)^2}{n_1} + \frac{(\sum M_2)^2}{n_2} + \text{.....} + \frac{(\sum M_5)^2}{n_5}\right\} - \frac{T^2}{N}$$

$$= \frac{(-12)^2}{5} + \frac{(2)^2}{5} + \frac{(-2)^2}{5} + \frac{(1)^2}{5} + \frac{(1)^2}{5} - 4 = 26.8$$

To find SSK, we arrange the data according to the letters row wise

A	B	C	D	E
6	-1	-5	-5	-7
6	-4	2	-2	0
4	-4	-4	2	0
8	0	-3	-4	0
12	-1	-1	-5	-4
36	-10	-11	-14	-11

$SSK = \frac{(36)^2}{5} + \frac{(-10)^2}{5} + \frac{(-11)^2}{5} + \frac{(-14)^2}{5} + \frac{(-11)^2}{5} - 4 = 362.8$ $SSE = SST - SSC - SSR - SSK$

$= 520 - 95.6 - 26.8 - 362.8 = 34.8$ $MSC = \frac{SSC}{n-1} = \frac{95.6}{4} = 23.9; MSR = \frac{SSR}{n-1} = \frac{26.8}{4} = 6.7;$

$MSK = \frac{SSK}{n-1} = \frac{362.8}{4} = 90.7; MSE = \frac{SSE}{(n-1)(n-2)} = \frac{34.8}{12} = 2.9$

$F_C = \frac{MSC}{MSE} = \frac{23.9}{2.9} = 8.24, F_R = \frac{MSR}{MSE} = \frac{6.7}{2.9} = 2.31$ $F_T = \frac{MSK}{MSE} = \frac{90.7}{2.9} = 31.28$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Between columns	SSC=95.6	4	MSC=23.9	F_C = 8.24
Between rows	SSR=26.8	4	MSR=6.7	F_R = 2.31
Between Treatments	SSK=362.8	4	MSK=90.7	F_T = 31.28
Residual (error)	SSE=34.8	12	MSE=2.9	
Total	SST=520			

Table value of F_C (4,12) at 1% level = 5.41

Table value of F_R (4,12) at 1% level = 5.41

Table value of F_T (4,12) at 1% level = 5.41

Inference : calculated value of F_C > the table value of F, H_0 is rejected at 1% level of significance.

∴ There is significant difference between columns.

calculated value of F_R < the table value of F, H_0 is accepted at 1% level of significance.

∴ There is no significant difference between rows.

calculated value of F_T > the table value of F, H_0 is rejected at 1% level of significance.

∴ There is significant difference between the treatments

5.5 2^2 Factorial Design

Factorial experiments involve simultaneously more than one factor and each factor is at two or more levels. Several factors affect simultaneously the characteristic under study in factorial experiments and the experimenter is interested in the main effects and the interaction effects among different factors.

Suppose in an experiment, the values of current and voltage in an experiment affect the rotation per minutes (rpm) of fan speed. Suppose there are two levels of current.

- 5 Ampere, call it as level 1 (C_1) and denote it as a_0
- 10 Ampere, call it as level 2 (C_1) and denote it as a_1.

Similarly, the two levels of voltage are

- 200 volts, call it as level 1 (V_0) and denote it as b_0
- 220 volts, call it as level 2 (V_1) and denote it as b_1.

The two factors are denoted as , A say for current and , B say for voltage.

In order to make an experiment, there are 4 different combinations of values of current and voltage.

1. Current = 5 Ampere and Voltage = 200 Volts, denoted as $C_0V_0 \equiv a_0b_0$
2. Current = 5 Ampere and Voltage = 220 Volts, denoted as $C_0V_1 \equiv a_0b_1$.
3. Current = 10 Ampere and Voltage = 200 Volts, denoted as $C_1V_0 \equiv a_1b_0$
4. Current = 10 Ampere and Voltage = 220 Volts, denoted as $C_1V_1 \equiv a_1b_1$

The responses from those treatment combinations are represented by $a_0b_0 \equiv (1), (a_0b_1) \equiv (b)$, $(a_1b_0) \equiv (a)$ and $(a_1b_1) \equiv (ab)$, respectively.

Now consider the following:

I. $\frac{(C_oV_o)+(C_oV_1)}{2}$: Average effect of voltage for the current level C_0

$$: \frac{(a_ob_o)+(a_ob_1)}{2} \equiv \frac{(1)+(b)}{2}$$

II. $\frac{(C_1V_o)+(C_1V_1)}{2}$: Average effect of voltage for the current level C_1

$$: \frac{(a_1b_o)+(a_1b_1)}{2} \equiv \frac{(a)+(ab)}{2}$$

Compare these two group means (or totals) as follows:

Average effect of V_1 level – Average effect at V_0 level

$$= \frac{(b)+(ab)}{2} - \frac{(1)+(a)}{2}$$

= Main effect of voltage

= Main effect of B.

Comparison like

(C_oV_1) - $(C_oV_o) \equiv (a)$ - (1): indicate the effect of voltage at current level C_o

and

(C_1V_1) - $(C_1V_0) \equiv (ab)$ - (b): indicate the effect of voltage at current level C_1.

The average interaction effect of voltage and current can be obtained as

$$\begin{pmatrix}\text{Average effect of voltage}\\ \text{at current level } I_o\end{pmatrix}-\begin{pmatrix}\text{Average effect of voltage}\\ \text{at current level } I_1\end{pmatrix}$$

$$= \text{Average effect of voltage at different levels of current.}$$

$$=\frac{(C_1V_1)-(C_1V_o)}{2}-\frac{(C_oV_1)-(C_oV_o)}{2}$$

$$=\frac{(ab)-(b)}{2}-\frac{(a)-(1)}{2}$$

$$=\text{Average interaction effect.}$$

Similarly

$$\frac{(C_oV_o)+(C_1V_o)}{2}=\frac{(1)+(b)}{2}:\text{Average effect of current at voltage level } V_0.$$

$$\frac{(C_oV_1)+(C_1V_1)}{2}=\frac{(a)+(ab)}{2}:\text{Average effect of current at voltage level } V_1$$

Comparison of these two as

$$\begin{pmatrix}\text{Average effect of current}\\ \text{at voltage level } V_0\end{pmatrix}-\begin{pmatrix}\text{Average effect of current}\\ \text{at voltage level } V_1\end{pmatrix}$$

$$=\frac{(C_0V_1)+(C_1V_1)}{2}-\frac{(C_0V_0)+(C_1V_0)}{2}$$

$$=\frac{(a)+(ab)}{2}-\frac{(1)+(b)}{2}$$

$$=\text{Main effect of current}$$

$$=\text{Main effect of } A.$$

Comparison like

$$(C_1V_0)-(C_0V_0)=(b)-(1):\text{Effect of current at voltage level } V_0$$

$$(C_1V_1)-(C_0V_1)=(ab)-(a):\text{Effect of current at voltage level } V_1:$$

The average interact effect of current and voltage can be obtained as

$$\begin{pmatrix}\text{Average effect of current}\\ \text{at voltage level } V_0\end{pmatrix}-\begin{pmatrix}\text{Average effect of current}\\ \text{at voltage level } V_1\end{pmatrix}$$

$$=\text{Average effect of current at different levels of voltage}$$

$$=\frac{(C_1V_1)-(C_0V_1)}{2}-\frac{(C_1V_0)-(C_0V_0)}{2}$$

$$=\frac{(ab)-(a)}{2}-\frac{(b)-(1)}{2}$$

$$=\text{Average interaction effect}$$

=Same as average effects of voltage at different levels of current.

(It is expected *that* the interaction effect of current and voltage is same as the interaction effect of voltage and current).

The quantity

$$\frac{(C_0V_0)+(C_1V_0)+(C_0V_1)+(C_1V_1)}{4}=\frac{(1)+(a)+(b)+(ab)}{4}$$

gives the **general mean effect** of all the treatment combination.

Treating (ab) as $(a)(b)$ symbolically (mathematically and conceptually, it is incorrect), we can now express all the main effects, interaction effect and general mean effect as follows:

$$\text{Main effect of } A=\frac{(a)+(ab)}{2}-\frac{(1)+(b)}{2}=\frac{1}{2}\left[(ab)-(b)+(a)-(1)\right]=\frac{(a-1)(b+1)}{2}$$

$$\text{Main effect of } B=\frac{(b)+(ab)}{2}-\frac{(1)+(a)}{2}=\frac{1}{2}\left[(ab)-(a)+(b)-(1)\right]=\frac{(a+1)(b-1)}{2}$$

$$\text{Interaction effect of } A \text{ and } B=\frac{(ab)-(b)}{2}-\frac{(a)-(1)}{2}=\frac{1}{2}\left[(ab)-(a)+(1)-(b)\right]=\frac{(a-1)(b-1)}{2}$$

$$\text{General mean effect } (M)=\frac{(1)+(a)+(b)+(ab)}{4}=\frac{1}{4}\left[(1)+(a)+(b)+(ab)\right]=\frac{(a+1)(b+1)}{4}$$

Notice the roles of + and – signs as well as the divisor.

- There are two effects related to A and B.
- To obtain the effect of a factor, write the corresponding factor with – sign and others with + sign. For example, in the main effect of A, a occurs with – sign as in $(a - 1)$ and b occurs with + sign as in $(b + 1)$.
- In AB, both the effects are present so a and b both occur with + signs as in $(a + 1)(b + 1)$.
- Also note that the main and interaction effects are obtained by considering the typical differences of averages, so they have divisor 2 whereas the general mean effect is based on all the treatment combinations and so it has divisor 4.
- There is a well defined statistical theory behind this logic but this logic helps in writing the final treatment combination easily. This is demonstrated later with appropriate reasoning.

Other popular notations of treatment combinations are as follows:

$a_0b_0 \equiv 0\ \ 0 \equiv I$
$a_0b_1 \equiv 0\ \ 1 \equiv a$
$a_1b_0 \equiv 1\ \ 0 \equiv b$
$a_1b_1 \equiv 1\ \ 1 \equiv ab.$

Sometimes 0 is referred to as 'low level' and 1 is referred to 'high level'.

Here I denote that both factors are at lower levels (a_0b_0 or 00). This is called as the **control treatment.**

These effects can be represented in the following table

Factorial effects	Treatment combinations				Divisor
	(1)	(a)	(b)	(ab)	
M	+	+	+	+	4
A	-	+	-	+	2
B	-	-	+	+	2
AB	+	-	-	+	2

The model corresponding to 2^2 factorial experiment is

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk},\ i = 1,2,\ j = 1,2,\ k = 1,2,...,n$$

where n observations are obtained for each treatment combinations.

When the experiments are conducted factor by factor, then much more resources are required in comparison to the factorial experiment. For example, if we conduct RBD for three-levels of voltage V_0, V_1 and V_2 and two levels of current I_0 and I_1, then to have 10 degrees of freedom for the error variance, we need

- 6 replications on voltage
- 11 replications on current.

So the total number of fans needed is 40.

For the factorial experiment with 6 combinations of 2 factors, the total number of fans needed are 18 for the same precision.

We have considered the situation up to now by assuming only one observation for each treatment combination, i.e., no replication. If r replicated observations for each of the treatment combinations are obtained, then the expressions for the main and interaction effects can be expressed as

$$A = \frac{1}{2r}\left[(ab) + (a) - b - (1)\right]$$

$$B = \frac{1}{2r}\left[(ab) + (b) - a - (1)\right]$$

$$AB = \frac{1}{2r}\left[(ab) + (1) - a - (b)\right]$$

$$M = \frac{1}{4r}\left[(ab) + (a) + (b) + (1)\right].$$

Let $Y_* = ((1),\ a,\ b,\ ab)'$ be the vector of total response values. Then

$$A = \frac{1}{2r}\ell_A' Y_* = \frac{1}{2r}(-1 \quad 1 \quad -1 \quad 1)Y_*$$
$$B = \frac{1}{2r}\ell_B' Y_* = \frac{1}{2r}(-1 \quad -1 \quad 1 \quad 1)Y_*$$
$$AB = \frac{1}{2r}\ell_{AB}' Y_* = \frac{1}{2r}(1 \quad -1 \quad -1 \quad 1)Y_*.$$

Note that A, B and AB are the linear contrasts. Recall that a linear parametric function is estimable only when it is in the form of linear contrast. Moreover, A, B and AB are the linear orthogonal contrasts in the total response values (1), , , a b ab except for the factor 1/2r.

The sum of squares of a linear parametric function $\ell' y$ is given by $\frac{(\ell' y)^2}{\ell'\ell}$. If there are r replicates, then the sum of squares is $\frac{(\ell' y)^2}{r\ell'\ell}$. It may also be recalled under the normality of y's, this sum of squares has a Chi-square distribution with one degree of freedom (χ_1^2). Thus the various associated sum of squares due to A, B and AB are given by the following:

$$SSA = \frac{(\ell_A' Y_*)^2}{r\ell_A'\ell_A} = \frac{1}{4r}(ab + a - b - (1))^2$$
$$SSB = \frac{(\ell_B' Y_*)^2}{r\ell_B'\ell_B} = \frac{1}{4r}(ab + b - a - (1))^2$$
$$SSAB = \frac{(\ell_{AB}' Y_*)^2}{r\ell_{AB}'\ell_{AB}} = \frac{1}{4r}(ab + (1) - a - b)^2.$$

Each of SSA, SSB and $SSAB$ has χ_1^2 under normality of Y_*.

The sum of squares due to total is computed as usual

$$TSS = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{r} y_{ijk}^2 - \frac{G^2}{4r}$$

where

$$G = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{r} y_{ijk}$$

is the grand total of all the observations.

The TSS has χ^2 distribution with $(2^2 r - 1)$ degrees of freedom.

The sum of squares due to error is also computed as usual as

$$SSE = TSS - SSA - SSB - SSAB$$

which has χ^2 distribution with

$$(4r-1)-1-1-1 = 4(r-1)$$

degrees of freedom.

The mean squares are

$$MSA = \frac{SSA}{1},$$

$$MSB = \frac{SSB}{1},$$

$$MSAB = \frac{SSAB}{1},$$

$$MSE = \frac{SSA}{4(r-1)}.$$

The F-statistic corresponding to A, B and AB are

$$F_A = \frac{MSA}{MSE} \sim F(1, 4(r-1)) \text{ under } H_0,$$

$$F_B = \frac{MSB}{MSE} \sim F(1, 4(r-1)) \text{ under } H_0,$$

$$F_{AB} = \frac{MSAB}{MSE} \sim F(1, 4(r-1)) \text{ under } H_0.$$

The ANOVA table is case of 2^2 factorial experiment is given as follows:

Source	Sum of squares	Degrees of freedom	Mean squares	F
A	SSA	1	MSA	$F_A = \frac{MSA}{MSE}$
B	SSB	1	MSB	$F_B = \frac{MSB}{MSE}$
AB	$SSAB$	1	$MSAB$	$F_{AB} = \frac{MSAB}{MSE}$
$Error$	SSE	$4(r-1)$	MSE	
Total	TSS	$4r-1$		

The decision rule is to reject the concerned null hypothesis when the value of the concerned F statistic

$$F_{\text{effect}} > F_{1-\alpha}(1, 4(r-1)).$$

Example 16 : The following data represents the results of a 2^2 factorial design with 2 factors and levels each with four replications. Analyses the data for response.

Treatment Combination	Replications			
	I	II	III	IV
(1)	12	12.3	11.8	11.6
a	12.8	12.6	13.7	11.8
b	11.5	11.9	12.6	15
ab	14.2	14.5	14.4	

Solution : Let A and B be the two factors. Here n = 4.

We code the data by subtracting 12 from each value. The coded data is

	x_1	x_2	x_3	x_4	Total	x_1^2	x_2^2	x_3^2	x_4^2
(1)	0	0.3	-0.2	-0.3	-0.3	0	0.09	0.04	0.09
a	0.8	0.6	1.7	0.4	5.1	0.64	0.36	2.89	0.16
b	-0.5	-0.1	0.6	2	-0.2	0.25	0.01	0.36	4
ab	2.2	2.5	2.4	-0.23	10.1	4.84	6.25	5.76	0.04
					14.7	5.73	6.71	9.05	13.2

H_0: All the mean effects are equal.

H_1: Not all equal.

N = 16; T = 14.7 ; $\frac{T^2}{N} = \frac{(14.7)^2}{16} = 13.5$

A contrast = a + ab –b – (1) = 5.1 + 10.1 – (- 0.2) – (- 0.3) = 15.7

B contrast = b + ab –a – (1) = - 0.2 + 10.1 – 5.1 – (- 0.3) = 5.1

AB contrast = (1) + ab –a – b = - 0.3 + 10.1 – 5.1 – (- 0.2) = 4.9

$$SSA = \frac{(Acontrast)^2}{4n} = \frac{(15.7)^2}{4 \times 4} = 15.41$$

$$SSB = \frac{(Bcontrast)^2}{4n} = \frac{(5.1)^2}{4 \times 4} = 1.63 \quad SSAB = \frac{(ABcontrast)^2}{4n} = \frac{(4.9)^2}{4 \times 4} = 1.5$$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3 + \sum x_4^2\right] - \frac{T^2}{N} = 5.73 + 6.71 + 9.05 + 13.2 - 13.5 = 21.2$$

$$SSE = SST - SSA - SSB - SSAB$$

$$= 21.2 - 15.41 - 1.63 - 1.5 = 2.66$$

$$MSA = SSA = 15.41; MSB = SSB = 1.63; MSAB = SSAB = 1.5$$

$$MSE = \frac{SSE}{4(n-1)} = \frac{2.66}{12} = 0.22$$

$$F_A = \frac{MSA}{MSE} = \frac{15.41}{0.22} = 69.41; F_B = \frac{MSB}{MSE} = \frac{1.63}{0.22} = 7.34$$

$$F_{AB} = \frac{MSAB}{MSE} = \frac{1.5}{0.22} = 6.76$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Factor A	SSA= 15.41	1	MSA = 15.41	F_A= 69.41
Factor B	SSB = 1.63	1	MSB = 1.63	F_B = 7.34
Interaction AB	SSAB = 1.5	1	MSAB = 1.5	F_{AB} = 6.76
Error	SSE = 2.66	12	MSE = 0.22	
Total	SST = 21.2			

Table value of F (1,12) at 5% level = 4.75

Inference :

Factor A : calculated value of F_A > the table value of F, H_0 is rejected at 5% level of significance.

∴ The effect of A is significant.

Factor B : calculated value of F_B > the table value of F, H_0 is rejected at 5% level of significance.

∴ The effect of B is significant.

Interaction AB : calculated value of F_{AB} > the table value of F, H_0 is rejected at 5% level of significance.

∴ The effect of interaction AB is significant.

Example 17 : The following data are obtained from a 2^2 factorial experiment replicated three times. Evaluate the sum of the squares for all factorial effect by the contrast method. Draw conclusions.

Treatment combination	Replicate I	Replicate II	Replicate III
(1)	12	19	10
a	15	20	16
b	24	16	17
ab	24	17	29

Solution : Let A and B be the two factors. Here n = 3.

We code the data by subtracting 20 from each value. The coded data is

	x_1	x_2	x_3	Total	x_1^2	x_2^2	x_3^2
(1)	-8	-1	-10	-19	64	1	100
a	-5	0	-4	-9	25	0	16
b	4	-4	-3	-3	16	16	9
ab	4	-3	9	10	16	9	49
				-21	121	26	174

H_0: All the mean effects are equal.

H_1: Not all equal.

N = 12; T = - 21 ; $\frac{T^2}{N} = \frac{(-21)^2}{12} = 36.75$

A contrast = a + ab –b – (1) = - 9 + 10 – (- 3) – (- 19) = 23

B contrast = b + ab –a – (1) = - 3 + 10 – (- 9) – (- 19) = 35

AB contrast = (1) + ab –a – b = - 19 + 10 – (- 9) – (- 3) = 3

$$SSA = \frac{(Acontrast)^2}{4n} = \frac{(23)^2}{4 \times 3} = 44.08$$

$$SSB = \frac{(Bcontrast)^2}{4n} = \frac{(35)^2}{4 \times 3} = 102.08 \quad SSAB = \frac{(ABcontrast)^2}{4n} = \frac{(3)^2}{4 \times 3} = 0.75$$

$$SST = \left[\sum x_1^2 + \sum x_2^2 + \sum x_3^3\right] - \frac{T^2}{N} = 121 + 26 + 174 - 36.75 = 284.25$$

$$SSE = SST - SSA - SSB - SSAB$$

$$= 284.25 - 40.08 - 102.08 - 0.75 = 141.34$$

$$MSA = SSA = 44.08; MSB = SSB = 102.08; MSAB = SSAB = 0.75$$

$$MSE = \frac{SSE}{4(n-1)} = \frac{141.34}{8} = 17.66$$

$$F_A = \frac{MSA}{MSE} = \frac{44.08}{17.66} = 2.49; F_B = \frac{MSB}{MSE} = \frac{102.08}{17.66} = 5.78$$

$$F_{AB} = \frac{MSE}{MSAB} = \frac{17.66}{00.75} = 23.54$$

ANOVA Table

Source of Variation	Sum of squares	D.F	Mean Square	Variance ratio
Factor A	SSA= 44.08	1	MSA = 44.08	F_A= 2.49
Factor B	SSB = 102.08	1	MSB = 102.08	F_B = 5.78
Interaction AB	SSAB = 0.75	1	MSAB = 0.75	F_{AB} = 23.54
Error	SSE = 141.34	8	MSE = 141.34	
Total	SST = 284.25			

Table value of F (1,8) at 5% level = 5.32

Table value of F (8,1) at 5% level = 4.75

Inference :

Factor A : calculated value of F_A < the table value of F, H_0 is accepted at 5% level of significance.

∴ The effect of A is not significant.

Factor B : calculated value of F_B > the table value of F, H_0 is rejected at 5% level of significance.

∴ The effect of B is significant.

Interaction AB : calculated value of F_{AB} < the table value of F(8,1), H_0 is accepted at 5% level of significance.

∴ The effect of interaction AB is not significant.

Summary

With two-way ANOVA, we are not only able to study the effect of two independent variables, but also the interaction between these variables. There are several advantages to conducting a two-way ANOVA, including efficiency, control of variables, and the ability to study the interaction between variables.

ANOVA includes calculating the following:

- Variation within the group (within-cell variation)
- Variation in the dependent variable attributed to one independent variable
- Variation in the dependent variable attributed to the other independent variable
- Variation between the independent variables

Keywords:

- ANOVA
- correction factor $\frac{T^2}{N}$

- SSC: Sum of squares between column
- SSR: Sum of squares between Row
- SST: Sum of squares between treatment
- MSC: Mean sum of squares between column
- MSR: Mean sum of squares between row
- MSE: Mean sum of Error
- Latin Square design: A Latin square design is a method of placing treatments so that they appear in a balanced fashion within a square block or field. Treatments appear once in each row and column. Replicates are also included in this design.
- Factorial experiments: involve simultaneously more than one factor and each factor is at two or more levels. Several factors affect simultaneously the characteristic under study in factorial experiments and the experimenter is interested in the main effects and the interaction effects among different factors.

SELF-ASSESSMENT QUESTIONS

Short Answer Questions:

1. What do you understand by Design of Experiments?
2. Distinguish between experimental variables and extraneous variables.
3. What is the aim of design of experiments?
4. State the basic principles of design of experiments.
5. What do you mean by experimental group and control group?
6. Name the basic designs of experiments.
7. Explain ANOVA.
8. Name the basic designs of experiments
9. Explain CRD.
10. Explain RBD.
11. Compare LSD and RBD
12. Write down the ANOVA table for one way classification.

Long Answer Questions

13. The following table shows the lives in hours of four brands of electric lamps. (Ans. F = 2.21, H_0 accepted)

A	1610	1610	1650	1680	1700	1720	1800	
B	1580	1640	1640	1700	1750			
C	1460	1550	1600	1620	1640	1660	1740	1820
D	1510	1520	1530	1600	1680			

Perform an analysis of variance test the homogeneity of the mean lives of the four brands of Lamps.

14. The accompanying data resulted from an experiment comparing the degree of soiling for fabric copolymerized with the three different mixtures of mathacrcylicacid. Analysis is the given classification.

Mixture 1	0.56	1.12	0.9	1.07	0.94
Mixture 2	0.72	0.69	0.87	0.78	0.91
Mixture 3	0.62	1.08	1.07	0.99	0.93

15. Four machines A, B, C,D are used to produce a certain kind of cotton fabric. Samples of size 4 with each unit as 100 square meters are selected from the outputs of the machines at random and the number of flaws in each 100 square meters are counted, with the following result.

A	8	9	11	12
B	6	8	10	4
C	14	12	18	9
D	20	22	25	23

Do you think that there is a significant difference in the performance of the four machines? (Ans. F = 25.21, H_0 rejected)

16. As head of a department of a consumers research organization you have the responsibility of testing and comparing lifetimes of four brands of electric bulbs. Suppose you test the lifetime of three electric bulbs each of 4 brands, the data is given below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours.

A	B	C	D
20	25	24	23
19	23	20	20
21	21	22	20

Can we infer that the mean lifetime of the four brands of electric bulbs are equal? (Ans. F = 25.21, H_0 accepted)

17. There are three main brands of a certain powder. A set of 120 sample values is examined and found to be allocated among four groups (A, B, C, D) and three brands (I, II, III) as shown hereunder.

	Groups			
Brands	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	8	19	11	13

Is there any significant difference in brands preference?

Answer at 5% level. (Ans. F = 3.69, H_0 rejected).

18. Perform two-way ANOVA for the given below:

	Treatment			
Plots of Land	A	B	C	D
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Ans. $F_C = 4.76$, $F_R = 5.14$, H_0 accepted)

19. The following table gives monthly sales (in thousand rupees) of a certain firm in the three states by its four sales men.

	Sales men			
States	I	II	III	IV
A	6	5	3	8
B	8	9	6	5
C	10	7	8	7

Setup the analysis of variance table and test whether there is any significant difference i) between sides by the firm salesmen and ii) between sales in the three states. (Ans. $F_C = 8.94$, $F_R = 5.14$, H_0 accepted)

20. Five doctors each test treatments for a certain disease and observe the number of days each takes to recover. The results are as follows (recovery time in days).

	Treatment				
Doctor	1	2	3	4	5
A	10	14	23	19	20
B	19	12	07	16	21
C	12	13	17	17	19
D	15	15	14	15	20
E	18	11	12	11	22

Carry out an analysis of variance and discuss the difference between i) doctors and ii) treatments. (Ans. $F_C = 47$, H_0 rejected; $F_R = 2.99$, H_0 accepted)

21. To study the performance of three detergents and three different water temperatures, the following whiteness readings were obtained with specially designed equipment.

Water Temperature	Detergent A	Detergent B	Detergent C
Cold Water	57	55	67
Warm Water	49	52	68
Hot Water	54	46	58

Perform a two-way analysis of variance, using 5% level of significance. (Ans. $F_C = 9.85$, H_0 rejected; $F_R = 2.38$, H_0 accepted)

22. Identify the following design, perform an analysis of variance for the design and comment on your findings.

C25	B23	A20	D20
A19	D19	C21	B18
B19	A14	D17	C20
D17	C20	B21	A15

(Ans. $F_C = 1.43$, H_0 accepted; $F_R = 8.86$, H_0 rejected;$F_T = 9.24$, H_0 rejected)

23. In a Latin square experiment given below are the yields in quintals per acre on the paddy crop carried out for testing the effect of five fertilizers A, B, C, D, E. Analyses the data for variations.

B25	A18	E27	D30	C27
A19	D31	C29	E26	B23
C28	B22	D33	A18	E27
E28	C26	A20	B25	D33
D32	E25	B23	C28	A20

(Ans. $F_C = 3.26$, H_0 rejected; $F_R = 1.22$, H_0 accepted; $F_T = 122$, H_0 rejected)

24. An agricultural experiment on the Latin square design gave the following results for the yield of wheat per acre, the letters corresponding to varieties, columns to treatments and rows to blocks. Discuss the variation of yield with each of these factors

A16	B10	C11	D9	E9
E10	C9	A14	B12	D11
B15	D8	E8	C10	A18
D12	E6	B13	A13	C12
C13	A11	D10	E7	B14

(Ans. $F_C = 37.82$, H_0 rejected; $F_R = 1.23$, H_0 accepted; $F_T = 69.64$, H_0 rejected)

25. Analyses the following results of a Latin square experiment.

A12	D20	C16	B10
D18	A14	B11	C14
B12	C15	D19	A13
C16	B11	A15	D20

(Ans. $F_C = 1.3$, H_0 accepted; $F_R = 1.08$, H_0 accepted; $F_T = 44.6$, H_0 rejected)

26. The following is a Latin square lay out of a design when 4 varieties of seeds are being tested set up the analysis of variance table and state your conclusion.

A70	B75	C68	D81
D66	A59	B55	C63
C59	D66	A39	B42
B41	C57	D39	A55

(Ans. $F_C = 3.2$, H_0 accepted; $F_R = 11.9$, H_0 rejected; $F_T = 2.02$, H_0 accepted)

27. Given the following observations for two factors A and B at two levels compute i) the main effects ii) make an analysis of variance

Treatment Combination	Replication I	Replication II	Replication III
(1)	10	14	9
a	21	19	23
b	17	15	16
ab	20	24	25

(Ans. $F_A = 45.63$, H_0 rejected; $F_B = 7.74$, H_0 rejected; $F_{AB} = 1.42$, H_0 accepted)

FURTHER READINGS

1. Devore, J.L, Probability and Statistics for Engineering and Sciences, Cengage Learning, 8th Edition, New Delhi, 2014.
2. Miller and M. Miller, Mathematical Statistics, Pearson Education Inc., Asia 7th Edition, New Delhi, 2011.
3. Richard Johnson, Miller and Freund's Probability and Statistics for Engineer, Prentice Hall of India Private Ltd., 8th Edition, New Delhi, 2011.
4. Jones, B. and M.G. Kenward. 1989. Design and analysis of crossover trials. Chapman and Hall, London.
5. Kuehl, R.O. 2000. Design of experiments: statistical principles in research design and analysis. Duxbury Press, Pacific Grove, CA.