THEORY OF LINEAR ESTIMATION AND ANALYSIS OF VARIANCE

M.Sc. STATISTICS

SEMESTER-II, PAPER-III

LESSON WRITERS

Prof. V.V. Haragopal
Professor of Statistics,
Department of Mathematics,
BITS Pilani, Jawaharnagar, Hyderabad

Dr. Bala Naga Hima Bindu Inampudi
Sr. Assistant Professor of Statistics
Department of Humanities and Services,
CVR College of Engineering, Hyderabad

Dr. B. Hari Mallikarjuna Reddy
Assistant Professor,
Dr. YSRHU-College of Horticulture,
Anantharajupeta,Annamayya District

Dr. M. Amulya
Assistant Professor,
Malla Reddy Deemed to be University,
Maisammaguda, Medchal, Malkajgiri Dt.

EDITOR

Prof. G.V.S.R. Anjaneyulu
Professor (Rtd.)
Department of Statistics,
Acharya Nagarjuna University

ACADEMIC ADVISOR

Prof. G.V.S.R. Anjaneyulu
Professor (Rtd.)
Department of Statistics,
Acharya Nagarjuna University

DIRECTOR, I/c.

Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W., M.Phil., Ph.D.
Professor

CENTRE FOR DISTANCE EDUCATION
ACHARYA NAGARJUNA UNIVERSITY
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863-2346259 (StudyMaterial)
Website www.anucde.info
E-mail:anucdedirector@gmail.com

M.Sc. STATISTICS: THEORY OF LINEAR ESTIMATION AND ANALYSIS OF VARIANCE

First Edition : 2025

No. of Copies :

© Acharya Nagarjuna University

This book is exclusively prepared for the use of students of M.Sc. Statistics,Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Publishedby:

Prof. V. VENKATESWARLU
Director, I/c
CentreforDistanceEducation,
AcharyaNagarjunaUniversity

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com. courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

M.Sc. STATISTICS
SEMESTER-II, PAPER-III
203ST24-THEORY OF LINEAR ESTIMATION AND ANALYSIS OF VARIANCE
SYLLABUS

UNIT-I:

Matrix algebra- Fundamental definitions, determinants, rank of a matrix, inverse of a matrix, orthogonal matrix, idempotent matrix, characteristic roots and vectors of a matrix. Numerical computation of characteristic roots and vectors for a positive definite matrix. Reduction of a positive definite matrix to a diagonal form using an Orthogonal matrix and non-singular matrix. Cayley-Hamilton theorem, trace of a matrix. Quadratic forms, reduction of quadratic forms using orthogonal transformation, statement of Cochran's theorem for quadratic forms.

UNIT-II:

Theory of linear estimation, linear models, estimability of linear parametric function, best linear unbiased estimator, Gauss-Markov set-up, Gauss-Markov theorem, generalized linear model, generalized Gauss-Markov theorem (Atken's theorem).

UNIT-III:

Decomposition of sum of squares in analysis of variance one-way classification, two-way classification with equal and unequal number of observations per cell. Multiple comparisons; Fisher's least significance difference test and Duncan's multiple range test, Fixed, random and mixed effect models.

UNIT-IV:

Analysis of covariance of one way and two-way classification, applications to standard designs- CRD, RBD missing plot technique- general theory and applications to RBD and LSD.

UNIT-V:

Model Adequacy checking: Test for Normality, Test for equality of Variances (Bartlett test, Modified Levene Method). Multiple comparison tests: Turkey's test, The Fisher Least Significant Difference (LSD) method, Duncan's Multiple range test.

BOOKS FOR STUDY:

1) Montgomery, D.C, (1976), Design and Analysis of Experiments., John Wiley &Sons.
2) Joshi, D.D.(1987), Linear Estimation and Design of Experiments., Wiley Eastern Ltd.
3) Das, M.N. and Giri, N.C. (1986), Design and An Analysis of Experiments, Wiley EasternLtd.

BOOKS FOR REFERENCES:

1) Datta, K.B. (2000), Matrix and Linear Algebra.

2) Rangaswamy, R, (1995), A Text Book of Agricultural Statistics, New Age International Publishers Limited.

3) Kempthorne, O, (1951), The design and Analysis of Experiments, Wiley Eastern Private Limited.

4) Rao, C.R, (1983), Linear Statistical Inference and its Applications, Wiley Eastern Ltd.

5) Raghavarao, D. (1987), statistical Techniques in Agricultural and Biological Research, Oxford&IBH Publishing Company Private limited.

6) Federer, Wt (1967), Experimental Design Theory and Application, Oxford & IBHPublishing Company.

7) Biswas, S. (1984). Topics in Algebra of Matrices, Academic Publication.

MODEL QUESTION PAPER (203ST24)

M.Sc. DEGREE EXAMINATION,
STATISTICS - SECOND SEMESTER
THEORY OF LINEAR ESTIMATION AND ANALYSIS OF VARIANCE

Time: Three Hours **Maximum: 70 Marks**

ANSWER ONE QUESTION FROM EACH UNIT

(Each question carries equal marks)

UNIT-I

1) a) Explain inverse matrix and idempotent matrix

b) State and prove Cayley-Hamilton theorem

2) a) Explain (i) determinant (ii) rank of a matrix and (iii) Inverse of a matrix with suitable example.

b) State and prove a necessary and sufficient condition for a real matrix to be positive definite

UNIT-II

3) a) Explain (i) linear model (ii) Best linear unbiased estimate.

b) State and prove Gauss – Markov theorem.

4) a) Describe Generalized linear model with suitable example.

b) State and prove Aitken’s theorem.

UNIT-III

5) a) Describe one-way classification for equal no. of observations per cell

b) Explain Duncan’ multiple range test

6) a) Explain Fisher’s least significant difference method

b) Explain analysis of variance two-way classification with multiple observations per cell.

UNIT-IV

7) a) Write the applications of CRD and RBD.

b) Explain analysis of covariance with a single concomitant variable.

8) a) Explain the analysis of variance two-way classification.

b) Explain analysis of LSD with one missing value.

UNIT-V

9) a) State and prove Bartlett’s test.

b) Briefly explain test for normality difference of variances.

10) a) What is multiple range test and its properties.

b) State and prove Turkey’s test.

CONTENTS

LESSON-1

FUNDAMENTALS OF MATRICES

1.0 OBJECTIVES:

After completing this lesson, students will be able to:

- Define and explain fundamental matrix concepts such as order, types of matrices, determinants, rank, and inverse.
- Compute determinants and rank using standard algebraic methods and elementary transformations.
- Determine the inverse of a matrix (when it exists) using adjoint and row-reduction techniques.
- Identify and verify properties of orthogonal and idempotent matrices, and understand their role in statistical models.
- Apply basic matrix operations and properties to solve simple linear algebra problems relevant to linear models and estimation.

STRUCTURE:

1.1 Introduction

1.2 Fundamental Matrix Definitions

1.3 Addition, Multiplication of Matrices

1.4 Properties

1.5 Conclusion

1.6 Self-Assessment Questions

1.7 Suggested Readings

1.1. INTRODUCTION:

Matrix algebra forms the backbone of many statistical techniques used in estimation, inference, and data analysis. In linear models, observations and parameters are conveniently expressed using matrices, allowing complex relationships to be handled in a compact and systematic manner. A strong foundation in matrix concepts is therefore essential for students of statistics and data science.

Determinants, rank, and inverse of matrices play an important role in solving systems of linear equations and in determining the existence and uniqueness of solutions. These ideas are particularly important in regression analysis and in deriving least squares estimators, where matrix operations simplify theoretical derivations.

Special matrices such as orthogonal and idempotent matrices occur naturally in the study of projection operators and sum of squares decomposition in ANOVA. Their algebraic properties help in understanding how variation in data is partitioned and how estimators behave under linear transformations.

Another important area in matrix algebra is the study of quadratic forms, which provides the mathematical framework for expressing many statistical quantities such as variance, sums of squares, and test statistics. Mastery of these concepts allows students to analyze multivariate data and to interpret geometrical aspects of statistical models.

In this lesson, we introduce the fundamental concepts of matrix algebra that support the theory of linear estimation. The aim is to equip students with the necessary tools to understand later topics such as diagonalization, quadratic forms, and Cochran’s theorem, which are central to advanced statistical inference.

1.2. FUNDAMENTAL MATRIX DEFINITIONS:

Definition of a Matrix:

‘A’ set of $m \times n$ numbers (scalars) arranged in the form of a rectangular array denoted by

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & & a_{in} \\ \vdots & \vdots & \vdots & & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & & a_{mn} \end{vmatrix}$$

Here ' m ' is number of rows and ' n ' is number of columns is called a matrix of order $m \times n$. The numbers $a_{11}, a_{12}, \dots\dots\dots .. a_{mn}$ are called the elements of the matrix. The matrix ' A ' can be represented by $A = \left[a_{ij}\right]_{m \times n}$, where $i = 1,2, \dots . m$ and $j = 1,2, \dots . n$.

In the matrix the horizontal lines are called rows (or) row. vectors & vertical lines are called columns (or) column vectors.

For example a_{34} is the element of intersection (or) 3^{rd} row & 4th column of the matrix. Here there are few examples of the matrix.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{2\times2} ; \; B = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}_{1\times3} ; \; C = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}_{3\times3}$$

$$D = \begin{bmatrix} 4 & 1 & 2 & 3 \\ 5 & 6 & 7 & 8 \\ 0 & 1 & 2 & 3 \\ 1 & 4 & 3 & 1 \end{bmatrix}_{4\times 4} ; E = \begin{bmatrix} a & b & c & d & e \\ f & g & h & i & j \\ k & l & m & n & o \\ p & q & r & s & t \\ u & v & \omega & x & y \end{bmatrix}_{5\times 5}$$

Row Matrix:

A matrix which has only one row is called Row Matrix. i.e., $1 \times n$ matrix.

Ex: - $[1 \quad 2 \quad 3]_{1\times 3};\quad [a \quad b \quad c]_{1\times 3}$

Column Matrix:

A matrix which has only one column is called column Matrix. i.e., $m \times 1$ matrix.

Ex:- $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}_{3\times 1}$ Note: $\begin{matrix} m = 3 \\ n = 1 \end{matrix}$

Square Matrix:

A matrix in which the no. of rows and no. of columns are equal then matrix is called Square Matrix.

Ex: $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{2\times 2} ; B = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}_{3\times 3}$

Rectangular Matrix:

A matrix in which the no. of rows is not equal to the no. of columns is called Rectangular Matrix.

Note $m \neq n$

Eg:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}_{2\times 4} ; B = \begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & 1 \end{bmatrix}_{3\times 4}$$

Determinant of a Matrix:

Let ' A ' be a square matrix the determinant of $'A'$ is the sum of the product of elements of any row or column with their co-factors it is denoted by $A(or)|A|$.

$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then, to write the determinant of ' A ' is $|A| = ad - bc$

Eg :

(i) $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \Rightarrow |A| = ad - bc = 4 - 6 = -2.$

(ii) $A = \begin{bmatrix} a & c \\ a & c \end{bmatrix} \Rightarrow |A| = ac - ac = 0$

Identity (or) Unit Matrix:

Square matrix each of whose diagonal element one and each of whose non-diagonal element is equal to zero is called unit matrix (or) identity matrix.

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

Null (or) Zero Matrix:

The matrix whose elements are all zeros is called a Null matrix (or) zero matrix.

$$0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Sub-matrix of a Matrix:

Any matrix obtained by ommitting some rows and columns from a given $m \times n$ matrix, is called a sub-matrix of a given matrix.

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 9 & 8 & 1 \end{bmatrix}_{3\times4} \; ; \; B = \begin{bmatrix} 2 & 3 \\ 6 & 7 \\ 9 & 8 \end{bmatrix}_{3\times2}$$

B is a sub-matrix of A.

Diagonal Matrix:

A Square matrix $A = [a_{ij}]_{n\times n}$ whose elements above and below the principal diagonal are zero is called a diagonal matrix.

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

Scalar Matrix:

A Diagonal matrix whose diagonal elements are equal is called a scalar Matrix.

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Inverse of a Matrix:

A Square matrix ' A ' of order $m \times n$ have the inverse matrix 'B' of order $n \times m$

if $AB = BA =$ I. If we write $B = A^{-1}$.

Eg:

(i) $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ & $|A| \neq 0$ [A is normal matrix]

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

$$A^{-1} = \frac{Adj(A)}{|A|} \left[\because \text{Adj}(A) = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}; |A| = ad - bc \right]$$

Ex:

1. $A = \begin{bmatrix} 1 & 2 \\ 3 & -5 \end{bmatrix}$ then to calculate Det$A$ & inverse of A?

$$A = \begin{bmatrix} 1 & 2 \\ 3 & -5 \end{bmatrix} \; |A| = ad - bc = -5 - 6 = -11 = 11 \neq 0$$

$$AdjA = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} -5 & -2 \\ -3 & 1 \end{bmatrix}$$

$$A^{-1} = \frac{Adj(A)}{|A|} = \frac{\begin{bmatrix} -5 & -2 \\ -3 & 1 \end{bmatrix}}{-11} = \begin{bmatrix} 5/11 & 2/11 \\ 3/11 & -1/11 \end{bmatrix}$$

$$\Rightarrow A = \begin{bmatrix} \cos a & -\sin a \\ \sin a & \cos a \end{bmatrix}$$

$$|A| = ad - bc = \cos^2 a + \sin^2 a = 1$$

$$Adj(A) = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} \cos a & \sin a \\ -\sin a & \cos a \end{bmatrix}.$$

$$A^{-1} = \frac{Adj(A)}{|A|} = \frac{\begin{bmatrix} \cos a & \sin a \\ -\sin a & \cos a \end{bmatrix}}{1} = \begin{bmatrix} \cos a & \sin a \\ -\sin a & \cos a \end{bmatrix}$$

Triangular Matrix:

(i) Upper Triangular Matrix:

A Square matrix $A = [a_{ij}]$ is called an upper triangular matrix. if $a_{ij} = 0$ whenever $i > j$

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

(ii) Lower - Triangular Matrix:

A Square matrix $A = [a_{ij}]$ is called a Lower -Triangular matrix. if $a_{ij} = 0$ whenever

$i < j$

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

Transpose Matrix:

Let $A = [a_{ij}]_{m \times n}$ then the $n \times m$ matrix obtained from A by changing its rows into columns and columns into rows is called Transpose of A and denoted by A^{T}.

Conjugate of a Matrix:

The matrix obtained from any given matrix A on replacing its elements by the corresponding conjugate complex number is called conjugate of A and denoted as A.

$$A = \begin{bmatrix} 2+3i & -2i \\ 4-6i & 3+i \end{bmatrix} \text{ then}$$
$$\bar{A} = \begin{bmatrix} 2-3i & 2i \\ 4+6i & 3-i \end{bmatrix}$$

Idempotent Matrix:

A square matrix ' A ' is said to be an Idempotent Matrix $A' = A$

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

$$A^2 = A \cdot A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{4}+\frac{1}{4} & \frac{1}{4}+\frac{1}{4} \\ \frac{1}{4}+\frac{1}{4} & \frac{1}{4}+\frac{1}{4} \end{bmatrix}$$
$$= \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$
$$A^2 = A$$

$\therefore A$ is an Idempotent Matrix.

Trace of a Matrix:

Let ' A ' be any square matrix. Then the sum of their principal diagonal elements of A is called Trace of a matrix. It is denoted by $\operatorname{tr}(A)$.

Eg: $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{2\times 2}$; $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$

Then $\operatorname{Trace}(A) = 1 + 4 = 5$ $\operatorname{Tr}(A) = 1 + 5 + 9 = 15.$

Symmetric Matrix:

A Square matrix $A = [a_{ij}]$ is said to be symmetric if $a_{ij} = a_{ji} \forall i,j$ A is symmetric if $A = A^{\top}$.

Eg:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}; \quad A^{\top} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$
$$\therefore A = A^{\top}$$

$$A = \begin{bmatrix} -2 & 3 & -4 \\ 3 & 1 & 0 \\ -4 & 0 & 3 \end{bmatrix}; \; A^{\mathsf{T}} = \begin{bmatrix} -2 & 3 & -4 \\ 3 & 1 & 0 \\ -4 & 0 & 3 \end{bmatrix}$$

$$\therefore A = A^{\mathsf{T}} \Rightarrow A \text{ is symmetric matrix.}$$

Skew-Symmetric Matrix:

A square matrix is called a skew-symmetric matrix if $A = (-A^{\mathsf{T}})$ (or) $A^{\mathsf{T}} = -A$. If $A = (a_{ij})_{\mathrm{m\times n}}$ is a skew-symmetric matrix. then $a_{ij} = -a_{ij}$ for all i & j. In a symmetric matrix each of the diagonal matrix is zero.

Eg:

$$A = \begin{bmatrix} 0 & 1 & -5 \\ -1 & 0 & 3 \\ 5 & -3 & 0 \end{bmatrix}, A^{\mathsf{T}} = \begin{bmatrix} 0 & -1 & 5 \\ 1 & 0 & -3 \\ -5 & 3 & 0 \end{bmatrix}$$

$$-A^{\mathsf{T}} = \begin{bmatrix} 0 & 1 & -5 \\ -1 & 0 & -3 \\ 5 & 3 & 0 \end{bmatrix}$$

$$A = -A^{\mathsf{T}}$$

A is a skew-symmetric matrix.

Real - Symmetric Matrix:

Let ' A ' be a $n \times n$ real symmetric matrix then there exists an orthogonal matrix P such that $P'AP = \Delta$ or, $A = P\Delta A'$ where Δ is a diagonal matrix.

Let

$$A = \begin{bmatrix} 3 & -4 \\ -4 & -3 \end{bmatrix}$$

$$P = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

$$\Delta = \begin{bmatrix} -5 & 0 \\ 0 & 5 \end{bmatrix}$$

$$\begin{aligned} \mathrm{A} &= P'\Delta P \\ &= \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} -5 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \\ &= \begin{bmatrix} 3 & -4 \\ -4 & -3 \end{bmatrix} \end{aligned}$$

Pair of Real - Symmetric Matrix:

let A and B be real $m \times m$ symmetric matrices of which B is P.d then ∃ a matrix ' R ' such that $A = [R^{-1}]^{-1}\Delta R^{-1}$ and $B = (R^{-1})^{-1}R^{-1}$ where Δ is a diagonal matrix.\

Orthogonal Matrix:

A Square matrix A is said to be Orthogonal if $A \times A^{\top} = A^{\top} \times A = I$, where I is a unit matrix.

$$A = \frac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{bmatrix} \text{ then } A^{\top} = \frac{1}{3}\begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix}$$

$$A \cdot A^{\top} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\therefore A \cdot A^{\top} = A^{\top} \cdot A = I$$

Hermitian Matrix:

A square matrix $A = \left[a_{ij}\right]$ is said to be Hermitial If $a_{ij} = \overline{a_{ji}}, \forall i, j$ that is $(i, j)^{\text{th}}$ elements of A is equal to conjugate complex of $(i, j)^{\text{th}}$ elements of A.

$$A = \begin{bmatrix} a & b + ci \\ b - ci & d \end{bmatrix}_{2\times 2.}$$

Skew - Hermitian Matrix:

A square matrix $A = \left[a_{ij}\right]$ is said to be skew Hermitian if $a_{ij} = -\bar{a}_{ji}\ \forall i, j$

Complex Matrix:

A square matrix $'A'$ is said to be unitary if $A^{\theta} \cdot A = A \cdot A^{\theta} = I$. where A^{θ} is the transpose of a conjugate of a complex matrix.

$A = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$ is a unitary matrix

Periodic Matrix:

A square matrix ' A ' is said to be periodic if there exists a positive integer k such that $A^{k+1} = A$, then k is called the Period of A.

Ex:- For idempotent matrix period of $A = 1$, because

$$A^2 = A \text{ i.e., } A^{1+1} = A$$

$$A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}; A^2 = A$$

Nilpotent Matrix:

A square matrix A is said to be nilpotent if ∃ a possible integer ' n ' such that $A^n = 0$ where 0 is null matrix.

Eg: $$A = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ 2 & -1 & -3 \end{bmatrix} \text{ then } A^2 = 0$$

A is nilpotent

Minor of a Matrix:

If A is an $m \times n$ matrix then the determinant of every submatrix of A is called a Minor of the matrix A .

$$A = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ 2 & -1 & -3 \end{bmatrix} \text{then} \begin{vmatrix} 1 & 3 \\ 2 & 6 \end{vmatrix} \text{is called 2-rowed minor of } A \text{ .}$$

Equal Matrix:

Two matrices A and B are said to be equal if they are of the same type and each element of one is equal to the corresponding elements of the order it is denoted by $A = B$.

Eg: If

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} ; B = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}$$

then $A = B \Leftrightarrow a = 1, b = 2, c = 3, d = -1$.

Elementary Matrix:

A matrix obtained from a unit-matrix by a single elementary transformation is called a Elementary Matrix.

Eg:

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

Involuntary Matrix:

A matrix ' A ' is said to be involuntary matrix if $A^2 = I$.

$$A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}; A^2 = A \cdot A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$
$$= \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} = I.$$

1.3. ADDITION, MULTIPLICATION OF MATRICES:

Addition of Matrix:

The Sum of two matrices ' A ' and ' B ' of same order is a new matrix denoted by $A + B$ whose elements are the sum of the corresponding elements of the two matrices A and B.

If A and B are two matrices of order $(\mathrm{m} \times \mathrm{n})$ then $A + B$ is also matrix of order. Therefore

Eg: $A = \begin{bmatrix} 1 & 5 \\ 6 & 3 \end{bmatrix} ; B = \begin{bmatrix} 7 & 2 \\ 0 & 4 \end{bmatrix}$

$$A + B = \begin{bmatrix} 1+7 & 5+2 \\ 6+0 & 3+4 \end{bmatrix} = \begin{bmatrix} 8 & 7 \\ 6 & 7 \end{bmatrix}$$

Multiplication of Matrices:

Two matrices A and B are compatible for multiplication only if the no. of columns of A is equal to the no. of rows of B.

Eg: $A = \begin{bmatrix} 1 & 2 \\ 15 & 6 \end{bmatrix}_{2\times 2}$ $B = \begin{bmatrix} 0 & 2 & 3 \\ 1 & 4 & 5 \end{bmatrix}_{2\times 3}$

Here, The no. of columns of $A = 2$

The no. of rows of $B = 2$

$A \times B$ it is possible to find AB.

$$A \times B = \begin{bmatrix} 1\times 0 + 2\times 1 & 1\times 2 + 2\times 4 & 1\times 3 + 2\times 5 \\ 15\times 0 + 6\times 1 & 15\times 2 + 6\times 4 & 15\times 3 + 6\times 5 \end{bmatrix} = \begin{bmatrix} 2 & 10 & 13 \\ 6 & 54 & 75 \end{bmatrix}$$

In general, the product of $A \times B$ is written as AB and is defined as

$$\begin{bmatrix} \text{First row of } Ax & 1^{\text{st}} \text{ row of } Ax & 1^{\text{st}} \text{ row of } Ax \\ \text{First column of } B & 2^{\text{nd}} \text{ column of } B & 3^{\text{rd}} \text{ column of } B \\ 2^{\text{nd}} \text{ row of } Ax & 2^{\text{nd}} \text{ row of } Ax & 2^{\text{nd}} \text{ row of } Ax \\ 1^{\text{st}} \text{ column of } B & 2^{\text{nd}} \text{ column of } B & 2^{\text{nd}} \text{ column of } B \end{bmatrix}$$

$$\begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{21} + a_{12}b_{22} & a_{11}b_{31} + a_{12}b_{32} \\ a_{21}b_{11} + a_{12}b_{12} & a_{21}b_{21} + a_{22}b_{22} & a_{21}b_{31} + a_{22}b_{32} \end{bmatrix}$$

1.4. PROPERTIES:

i) $AB \neq BA$

ii) $(AB)C = A(BC)$ [Assosciative law]

iii) $AI = IA =$ A [Existence of multiplicative identily.]

iv) For 2 matrices A_1B if $A \times B = 0$ it is not necessary that $A = 0$ (or) $B = 0$ (or) Both A and B are 'O'

Eg: If $A = \begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix}$ $B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, then

$$AB = \begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}_{2\times 2} = 0$$

v) $A(B + C) = (A \times B) + (A \times C)$ [left distributive law]

vi) $(A + B)C = AC + BC$ [Right distributive law]

1.5. CONCLUSION:

In this lesson, the foundational ideas of matrix algebra that form the basis for many statistical and computational techniques were introduced. Beginning with essential matrix definitions, students gained an understanding of how data and linear relationships can be represented compactly using rows, columns, and different types of matrices. These fundamental concepts prepare us for more advanced topics in linear models, multivariate analysis, and estimation theory.

Also, here explored basic matrix operations such as addition and multiplication, which are crucial for expressing and solving systems of linear equations. Understanding how matrices interact under these operations helps students to interpret model structures like ($X\beta$) and transformations such as ($A'XA$). These operations also reveal important algebraic behaviours - such as non-commutativity - that significantly influence statistical procedures and matrix decompositions.

Finally, here discussed selected matrix properties that frequently appear in statistical applications. Properties such as symmetry, diagonal dominance, invertibility, and orthogonality play a vital role in simplifying computations and understanding the geometry of linear transformations. Together, the concepts covered in this lesson provide the mathematical groundwork needed for studying determinants, rank, inverse matrices, and matrix transformations in the subsequent lessons.

1.6. SELF-ASSESSMENT QUESTIONS:

1) Describe in detail the importance of learning matrix algebra for data science, machine learning, and computational statistics. Give suitable examples.
2) Define and explain different types of matrices: row matrix, column matrix, square matrix, diagonal matrix, scalar matrix, identity matrix, and zero matrix. Illustrate each with examples.
3) State and prove the properties of matrix multiplication. Explain why matrix multiplication is not commutative, giving suitable examples.
4) Discuss in detail the associative, commutative, and distributive properties related to matrix addition and multiplication. Provide proofs and examples for each property.
5) Discuss the role of special matrices (identity, zero, diagonal matrices) in matrix operations. Explain how these matrices behave under addition, multiplication, and transposition.

1.7. SUGGESTED READINGS:

1) **Introduction to Linear and Matrix Algebra** – by Richard Bronson & Gabriel B. Costa
2) **Matrices and Linear Algebra** – by I. N. Herstein & D. J. Winter
3) **Linear Algebra and Matrices** – by K. Hari Kishan
4) **Linear Algebra and Matrices: Topics for a Second Course** – by Helene Shapiro

Dr. Bala Naga Hima Bindu, Inampudi.

LESSON-2

RANK OF A MATRIX

2.0. OBJECTIVES:

After completing this lesson, students will be able to:

- Understand the concept of rank
- Compute the rank of a matrix
- Determine the solvability of linear systems
- Analyze linear dependence and independence
- Apply the concept of rank.

STRUCTURE

2.1 INTRODUCTION

The rank of a matrix is one of the fundamental concepts in linear algebra and matrix theory. It essentially measures the "non-degenerateness" of a matrix by indicating the maximum number of linearly independent rows or columns it contains. In simpler terms, the rank tells us how much information a matrix carries and whether its rows or columns are redundant. It plays a central role in solving systems of linear equations, determining invertibility, and analyzing the dimensions of vector spaces associated with the matrix.

Mathematically, the rank of a matrix $A$is defined as the dimension of the row space or the column space of A. The row space is the set of all possible linear combinations of the rows of the matrix, and the column space is the set of all linear combinations of its columns. Interestingly, the row rank and column rank of any matrix are always equal, and this common value is referred to simply as the rank of the matrix. This property highlights the intrinsic symmetry in linear algebra between rows and columns.

The concept of rank is also closely linked to the idea of solutions to linear systems. If a system of linear equations is represented in matrix form as $AX = B$, the rank of the coefficient matrix Aprovides crucial information about the existence and uniqueness of solutions. Specifically, if the rank of Aequals the number of unknowns, the system has a unique solution; if it is less, the system may have infinitely many solutions or none, depending on the rank of the augmented matrix.

Finally, rank is widely used in various applications beyond solving equations. In statistics, it helps in determining the independence of variables, while in engineering and computer science, it assists in analyzing networks, transformations, and data structures. Understanding the rank of a matrix equips us with a tool to probe the structural properties of matrices and to address practical problems efficiently.

2.2. RANK OF A MATRIX:

Determinant:

The determinant is a scalar value that can be computed from the elements of a square matrix and encodes certain properties of the linear transformation described by the matrix. The determinant of a matrix A is denoted by $|A|$ or $\det A$.

Eg:

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

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

$$|A| = 10 - 21 = -11$$

Rank of a Matrix:

Suppose A is a non-zero matrix, a positive integer r is said to be the rank of A. If

i) $\exists\ a$ non-zero r-rowed minor of A.

ii) Every (rH) rowed r-rowed minor of A.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{3\times3}$$

In other words, let A be any non-zero matrix then the rank of a matrix is defined as the order of non-singular sub-matrix of A. It is denoted by rank (A)

Eg:- Let $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 5 \\ 3 & 4 & 2 \end{bmatrix}_{3\times3}$

$$\begin{aligned} \det|A| &= |1(2-20) - 2(4-15) + 3(8-3)| \\ &= |-18 + 22 + 15| \\ &= 19 \neq 0 \end{aligned}$$

$$\det A \neq 0$$

$\therefore$ A is non-singular matrix.

$\therefore$ So, Rank of (A) is 3

Singular Matrix:

A square matrix $'A'$ is said to be a singular Matrix if $|A| = 0$

Eg: $$A = \begin{bmatrix} 1 & 0 \\ 5 & 0 \end{bmatrix}$$

$|A| = 0$

Non-Singular Matrix:

A Square matrix ' A ' is said to be a non-singular matrix if $|A| \neq 0$.

Eg: $$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$|A| = 1$$

Multiplicative Inverse of a Square matrix:

Let $ad - bc = 0$, then the inverse of $'A'$ denoted byA^{-1} as defined as

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

1. If ad $-bc = 0, |A| = 0,$ i.e, ' A ' is a singular matrix then A^{-1} is not defined

2. $A \times A^{-1} = A^{-1} \times A = I.$

   Eg: If $A = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}$ then find A^{-1}

$|A| = 12 - 15 = -3$

$$A^{-1} = \frac{1}{-3}\begin{bmatrix} 6 & -3 \\ -5 & 2 \end{bmatrix} = \begin{bmatrix} 6/-3 & -3/-3 \\ -5/-3 & 2/-3 \end{bmatrix}$$
$$= \begin{bmatrix} -2 & 1 \\ 5/3 & -2/3 \end{bmatrix}$$

Methods to find Inverse of a Matrix:

1) Matrix - Inversion Method
2) Cramer's Method.

1) Matrix - Inversion Method:

In this method, we first express the given coefficient matrix ' x ' is called the variables matrix and ' B ' is called the constant matrix.

Eg:

$$ax + by = p$$
$$cx + dy = q$$
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} p \\ q \end{bmatrix}$$

If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, x = \begin{bmatrix} x \\ y \end{bmatrix}; B = \begin{bmatrix} p \\ q \end{bmatrix}$ then $Ax = \mathrm{B}$

multiplying this by A^{-1} on B is

$$A^{-1}(Ax) = A^{-1}B$$
$$Ix = A^{-1}B$$
$$x = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} p \\ q \end{bmatrix} \because [(ad - bc) \neq 0]$$
$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} dp & -bq \\ -cp & +aq \end{bmatrix}$$
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (dp - bq)/(ad - bc) \\ (-cp + qa)/(ad - bc) \end{bmatrix}$$
$$x = \frac{dp - bq}{ad - bc}; y = \frac{aq - cp}{ad - bc}$$

2) Cramer's Method:

Consider two linear equations $ax + by = p$ and $cx + dy = q$ expressing these equations in matrix equation form.

$$A = \begin{bmatrix} a & d \\ b & c \end{bmatrix} x = \begin{bmatrix} x \\ y \end{bmatrix}, B = \begin{bmatrix} p \\ q \end{bmatrix}$$

Then $|A| = ad - bc \neq 0$

Let $B_1 = \begin{bmatrix} p & b \\ q & d \end{bmatrix} B_2 = \begin{bmatrix} a & p \\ c & q \end{bmatrix}$

Now

$$x = \frac{|B_1|}{|A|}, \qquad y = \frac{|B_2|}{|A|}$$
$$x = \frac{dp - qb}{ad - bc}, \quad y = \frac{aq - cp}{ad - bc}$$

Orthogonal Matrix:

A square matrix A is said to be orthogonal if $A \times A^{\mathsf{T}} = A^{\mathsf{T}} \times A = I$ where I is a unit matrix.

$$A = \frac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{bmatrix} \text{ then } A^{\mathsf{T}} = \frac{1}{3}\begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix}$$

$$AA^{\mathsf{T}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_{3\times 3}$$
$$\therefore AA^{\mathsf{T}} = A^{\mathsf{T}}A = I.$$

Determinant of a Matrix:

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

$$\begin{aligned} \det A &= [2(24-12) - 3(30-6) - 1(10-4)] \\ &= 24 - 72 - 6 \\ &= -54 \neq 0 \end{aligned}$$

Properties:

i) $|A^{\mathsf{T}}| = |A|$ $\quad$ $\text{Adj}A = |A| \cdot I$

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

$|A^{-1}| = \frac{1}{|A|}$

If ' A ' is invertible, where $|A| \neq 0$ (or) A is non-singular matrix.

Adjoint of a Matrix:

Let A be a square matrix, then the transpose of the cofactor matrix of A is called Adjoint Matrix of A. It is denoled by Adj A.

i.e.,

$\text{Adj}A = [\text{ co-factor} \quad of \quad A]^{-1}$

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

cofactor of $A = \begin{bmatrix} +9 & -1 & +(-3) \\ +3 & +1 & -1 \\ -11 & -1 & +5 \end{bmatrix}$

$\text{adj}A = \begin{bmatrix} 9 & 3 & -11 \\ -1 & 1 & -1 \\ -3 & -1 & 5 \end{bmatrix}$

Vector of a Matrix:

If a matrix has only one row (or) one column it is called a vector

- A matrix having only one row is called Row vector. Eg: $[2 \quad 3 \quad 4]$
- A matrix having only one column is called column vector: Eg: $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$

2.3. VECTOR SPACE:

Let V be a non-empty set whose elements are called vector. F be a field whose elements are called Scalars. Then V is said to be Vector space over a field ' F '. It is denoted by $V(F)$ if

i) $(V,+)$ is an abelian group.

ii) ' V ' is closed under scalar multiplication $\forall a+F,\alpha\in V\exists\, a\alpha\in V$

iii) Scalar properties:

a) $a(\alpha+\beta)=a\alpha+a\beta;\forall;a\in F;\alpha,\beta\in V.$

b) $(a+b)\alpha=a\alpha+b\alpha,a,b\in F,\alpha\in V.$

c) $(ab)\alpha=a(b\alpha)\;\forall\; a,b\in F,\alpha\in V.$

d) $1\cdot\alpha=\alpha\cdot 1=\alpha\,,\forall\alpha\in V.$

Dimension of Vector Space:

Let $V(F)$ be a vector space, let ' S ' be a basis for $V(F)$ then the number of vectors in the basis

'S' is called as dimension of a vector space. It is denoted by "dim V ".

Eg:

Let 'V' be a vector space of all ordered pairs of R then $S=\{(1{,}0)(0{,}1)\}$ is basis for' V '.

$\dim V=$ The number of elements in the basis $=2$

Properties:

1) If the dimension of vector space is finite, then the vector space is called Finite dimension vector space.

2) If the dimension of vector space is infinite, then the vector space is called Infinite dimension vector space.

3) Any two basis of a vector space have same number of vectors.

4) Let $V(F)$ be a vector space W_1,W_2 are two subspaces then

$$\dim(W_1+W_2)=\dim W_1+\dim W_2=\dim(W_1+W_2)$$

5) Let $V(F)$ be a vector space ' W ' be a subspace of ' V ' then

$$\dim(v/w)=\dim v-\dim w.$$

Linearly Independent:

Let $V(F)$ be a vector space and $S=\{\alpha_1,\alpha_1,\dots\alpha_n\}$ be a non-emply subset of v to ' s ' is said to be linearly independent then there exists scalars $a_1\alpha_1+a_2\alpha_2+a_3\alpha_3+\cdots\cdots+a_n\alpha_n=0.$

Let $a_1=a_2=a_3=\cdots\ldots=0$

Linearly Dependent:

Let $V(F)$ be a vector space and $\mathrm{S} = \{\alpha_1, \alpha_2, \dots, \dots, \alpha_{\mathrm{n}}\}$ be a non-emply subset of ' v ' then ' s ' is said to be Linearly dependent then ∃ scalars.

$$a_1\alpha_1 + a_2\alpha_2 + \cdots \cdots \cdot + a_n\alpha_n = 0$$

let $a_1 = a_1 = a_3 = \cdots \ldots = a_n$ not all zeros

Basis:

Let $V(F)$ be a vectos space. Let ' S ' be a finile subset of ' V ' then ' S ' is said to be Basis for $V(F)$ if

(i) S is linearly independent (L.I)

(ii) 'S' spans 'V'.

Index of a Matrix:

Index of matrix A is defined as the no. of positive terms in the C form (or) natural form of a matrix. It is denoted by ' P '.

Let

$$A = \begin{bmatrix} 1 & -2 & +4 \\ -2 & 2 & 0 \\ 4 & 0 & -7 \end{bmatrix}$$

$$\begin{array}{l} R_2 \longrightarrow R_2 + 2R_1 \\ R_3 \rightarrow R_3 - 4R_1 \end{array} A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & -2 & 8 \\ 0 & 8 & -23 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & -2 & 8 \\ 0 & 8 & -23 \end{bmatrix}$$

$$\begin{array}{l} C_2 \longrightarrow C_2 + 2C_1 \\ C_3 \rightarrow C_3 - 4C_1 \end{array}$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 8 \\ 0 & 0 & 9 \end{bmatrix} \quad R_3 \rightarrow R_3 + 4R_2$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 9 \end{bmatrix} C_3 \rightarrow C_3 + 4C_2$$

Eigen values $= 1, -2, 9$

Index $(p) = 2, \mathrm{rank}(r) = 3, \mathrm{signature}(s) = 2p - r = 4 - 3 = 1$

2.4. PROBLEMS:

1) Find the rank of a matrix $A = \begin{bmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & -7 \end{bmatrix}$

Sol: Given $A = \begin{bmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & -7 \end{bmatrix}$

$R_1 \leftrightarrow R_2$

$$A = \begin{bmatrix} 1 & -1 & -2 & -4 \\ 2 & 3 & -1 & -1 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & -7 \end{bmatrix}$$

$R_2 \rightarrow R_2 - 2R_1; R_3 \rightarrow R_3 - 3R_1; R_4 \rightarrow R_4 - 6R_1$

$$\begin{bmatrix} 1 & -1 & -2 & -4 \\ 0 & 5 & 3 & 7 \\ 0 & 4 & 9 & 10 \\ 0 & 9 & 12 & 17 \end{bmatrix}$$

$C_2 \rightarrow C_2 + C_1; C_3 \rightarrow C_3 + 2C_1; C_4 \rightarrow C_4 + 4C_1$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 5 & 3 & 7 \\ 0 & 4 & 9 & 10 \\ 0 & 9 & 12 & 17 \end{bmatrix}$$

$R_2 \rightarrow R_2 - R_3$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -6 & -3 \\ 0 & 4 & 9 & 10 \\ 0 & 9 & 12 & 17 \end{bmatrix}$$

$R_3 \rightarrow R_3 - 4R_2;\ R_4 \rightarrow R_4 - 9R_2$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -6 & -3 \\ 0 & 0 & 33 & 22 \\ 0 & 0 & 66 & 44 \end{bmatrix}$$

$C_3 \rightarrow {}^{C_3}/_{33};\ C_4 \rightarrow {}^{C_4}/_{22}$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{-2}{11} & \frac{-3}{22} \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 2 \end{bmatrix}$$

$C_3 \rightarrow C_3 + \frac{2}{11}C_2;\ C_4 \rightarrow C_4 + \frac{3}{22}C_2$

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&1\\0&0&2&2\end{bmatrix} R_4 \to R_4 - 2R_3;\ C_4 \to C_4 - C_3$$

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix} \sim \begin{bmatrix}I_3&0\\0&0\end{bmatrix}$$

The rank is 3.

2) Find the rank of the matrix $A = \begin{bmatrix}1&2&-4&5\\2&-1&3&6\\6&1&9&7\end{bmatrix}$

Sol): Given $A = \begin{bmatrix}1&2&-4&5\\2&-1&3&6\\8&1&9&7\end{bmatrix}$

$$R_2 \to R_2 - 2R_1\,;\ R_3 \to R_3 - 8R_1$$

$$A = \begin{bmatrix}1&2&-4&5\\0&-5&11&-4\\0&-15&41&-33\end{bmatrix}$$

$$C_2 \to C_2 - 2C_1; C_3 \to C_3 + 4C_1; C_4 \to C_4 - 5C_1$$

$$A = \begin{bmatrix}1&0&0&0\\0&-5&11&-4\\0&-15&41&-33\end{bmatrix} \quad C_2 \to \frac{C_2}{-5}$$

$$A = \begin{bmatrix}1&0&0&0\\0&1&11&-4\\0&3&41&-33\end{bmatrix}$$

$$R_3 \to R_3 - 3R_2$$

$$A = \begin{bmatrix}1&0&0&0\\0&1&11&-4\\0&0&8&-21\end{bmatrix}$$

$$C_3 \to C_3 - 11C_2; C_4 \to C_4 + 4C_2$$

$$A = \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&8&-21\end{bmatrix}$$

$$C_3 \to C_3/8; C_4 \to C_4/-21$$

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

$$C_4 \to C_4 - C_3$$

$$A = \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{bmatrix} \sim [I_3 0]$$

Rank of $A = P(A) = 3$.

2.5. SOME IMPORTANT RESULTS:

1. Idempotent Matrix:

Prove that the sum of two idempotent matrix is also idempotent.

Proof: Given that $AB = BA = 0$

Let A, B are two idempotent matrices.

$A^2 = A; B^2 = B.$

To prove $A + B$ is idempotent.

$(A + B)^2 = A + B$

Now

$$\begin{aligned}(A + B)^2 &= (A + B)(A + B)\\ &= A^2 + AB + AB + B^2\\ (A + B)^2 &= A + 0 + 0 + B \; [\because A^2 = A; B^2 = B]\\ (A + B)^2 &= A + B\end{aligned}$$

$(A + B)$ is idempotent.

2) If A, B are independent and cumulative then prove that AB is idempotent.

Proof: Given A, B are idempotent

$A^2 = A\,; B^2 = B.$

Given A, B are cumulative

$AB = BA.$

To prove ' AB ' is idempotent

$(AB)^2 = AB$

Now $(AB)^2 = (AB)(AB)$

$$\begin{aligned}&= A(BA)\cdot B = A(AB)\cdot B\\ (AB)^2 &= (AA)(BB)\\ (AB)^2 &= A^2 \cdot B^2 = AB\\ (AB)^2 &= AB\end{aligned}$$

$\therefore$ ' AB ' is idempotent matrix.

Problems:

1) Orthogonal Matrix:

Prove that $A = \begin{bmatrix} -1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & -1/2 \end{bmatrix}$ is orthogonal.

Sol) Given that,

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

$$A^T = \begin{bmatrix} -1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & -1/2 \end{bmatrix}$$

$$A \cdot A^T = \begin{bmatrix} -1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} -1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & -1/2 \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4} & \frac{-1}{4}-\frac{1}{4}+\frac{1}{4}+\frac{1}{4} & \frac{-1}{4}+\frac{1}{4}-\frac{1}{4}+\frac{1}{4} & \frac{-1}{4}+\frac{1}{4}+\frac{1}{4}-\frac{1}{4} \\ \frac{-1}{4}-\frac{1}{4}+\frac{1}{4}+\frac{1}{4} & \frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4} & \frac{1}{4}-\frac{1}{4}-\frac{1}{4}+\frac{1}{4} & \frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4} \\ \frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4} & \frac{1}{4}-\frac{1}{4}-\frac{1}{4}+\frac{1}{4} & \frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4} & \frac{1}{4}+\frac{1}{4}-\frac{1}{4}-\frac{1}{4} \\ \frac{-1}{4}+\frac{1}{4}+\frac{1}{4}-\frac{1}{4} & \frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4} & \frac{1}{4}+\frac{1}{4}-\frac{1}{4}-\frac{1}{4} & \frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4} \end{bmatrix}$$

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

$$A \cdot A^T = I$$

' A ' is orthogonal matrix.

2. Linearly Independent:

1) Prove that $\{(2,3,4)(0,1,2)(-1,1,-1)\}$ are linearly independent.

Sol): Let $s = \{(2,3,4)(0,1,2)(-1,1,-1)\}$.

$$\Rightarrow a(2,3,4)+b(0,1,2)+c(-1,1,-1)=0$$
$$\Rightarrow (2a,3a,4a)+(0,b,2b)+(-c,c,-c)=0$$
$$(2a+0-c,3a+b+c,4a+2b-c)=0$$

comparing the LHS and RHS

$$2a+0-c=0 \quad ;3a+b+c=0;4a+2b-c=0$$
$$2a-c=0 \qquad \longrightarrow (2) \qquad \longrightarrow (3)$$
$$2a=c\longrightarrow(1)$$

$$(2)+(3)\Rightarrow \begin{array}{l} 4a+2b-c=0 \\ \underline{3a+b+c \quad =0} \\ 7a+3b \qquad =0 \end{array}$$

$$\longrightarrow (\mathrm{a})$$

from $\varepsilon q^n(2)$;

$$\begin{array}{r} 3a+b+c \qquad =0 \\ 3a+b+2a \qquad =0 \\ 5a+b=0 \end{array}$$

$$\longrightarrow (\mathrm{b})$$

(a) - 3(b) ⇒

$$\begin{array}{l} 7a+3b \qquad =0 \\ \underline{\pm15a\pm3b \quad =0} \\ -8a \qquad =0 \end{array}$$

$$a=0$$

Sub $a=0$ in eqn (1)

$$2a=c$$
$$c=0$$

Sub $a=0$ & $c=0$ in eq^n (2)

$$3(0)+b+0=0$$
$$b=0$$

$$\therefore a=b=c=0$$

' s ' is linearly independent

2.6. CONCLUSION:

In conclusion, determinants, rank, and linear dependence are closely interconnected concepts in matrix theory. The determinant of a square matrix provides a quick test of invertibility: a matrix is non-singular if its determinant is non-zero and singular if the determinant is zero. The rank of a matrix measures the maximum number of linearly independent rows or columns, indicating the "information content" of the matrix and determining the existence and uniqueness of solutions to linear systems. Meanwhile, linear independence of rows or columns ensures full rank, whereas linear dependence implies redundancy and reduces the rank. Together, these concepts form the foundation for understanding matrix behavior, solving linear equations, and analyzing vector spaces in both theoretical and applied contexts.

2.7. SELF-ASSESSMENT QUESTIONS:

1) Explain the method of determining the rank of a matrix using minors.
2) State and explain the properties of the rank of a matrix.
3) Explain the connection between rank and linear independence of rows/columns.
4) Show that if a matrix A has rank r, then A can be expressed as the product of two matrices-one of size $m \times r$and the other of size $r \times n$.

2.8. SUGGESTED READINGS:

1) **Introduction to Matrix Theory** - Arindama Singh
2) **Linear Algebra** - Jörg Liesen & Volker Mehrmann
3) **Matrix Theory and Linear Algebra** - Peter Selinge
4) **A Textbook of Matrices** - Hari Kishan
5) **Linear Algebra and Matrix Theory** - Robert R. Stoll

Dr. Bala Naga Hima Bindu, Inampudi

LESSON-3

QUADRATIC FORMS

3.0. OBJECTIVES:

After completing this lesson, students will be able to:

- Understand and explain the Cayley-Hamilton theorem including its statement, meaning, and importance in matrix theory and linear algebra.
- Apply the Cayley-Hamilton theorem to compute powers of matrices, find matrix inverses (when they exist), and simplify polynomial expressions in matrices.
- Identify and classify quadratic forms, and express them in matrix notation to analyze their structure.
- Perform reduction of quadratic forms to canonical form or diagonal form using orthogonal transformations or congruence transformations.
- Determine the nature of quadratic forms (positive definite, negative definite, indefinite, etc.) using eigenvalues, principal minors, and other criteria.
- Understand the statement and implications of Cochran's theorem in the context of quadratic forms and sums of squares in statistics.
- Integrate the concepts of matrix algebra (Cayley-Hamilton theorem), quadratic forms, and Cochran's theorem to solve advanced problems in linear algebra, multivariate analysis, and statistical inference.

STRUCTURE

3.1. INTRODUCTION

The Cayley-Hamilton theorem, quadratic forms, and Cochran's theorem are important tools that help students understand deeper structures in matrix theory and its applications. The Cayley-Hamilton theorem is significant because it allows complex matrix calculations to be simplified using the matrix's own characteristic equation. This makes it easier to compute matrix powers, understand matrix behaviour, and solve systems that involve repeated transformations. It has applications in engineering, computer science, control theory, and any area where linear systems evolve over time.

Quadratic forms and Cochran's theorem are widely used in geometry, optimization, economics, and especially in statistics. Quadratic forms help us to classify surfaces, study the nature of functions, and determine whether a system is stable or unstable. They are also essential in statistical methods such as least squares, regression, and multivariate analysis. Cochran's theorem adds further importance by explaining how total variation in statistical models can be broken into meaningful components, which is the foundation of ANOVA and variance estimation. Together, these topics build strong analytical skills and provide practical methods for solving real-world problems in science, data analysis, and applied mathematics.

3.2. CHARACTERISTIC EQUATION, CAYLEY -HAMILTON THEOREM:

Characteristic Equation:

Let ' A ' be a square matrix. $|A|$ is determinant then $|A-\lambda I|=0$ is called as Characteristic Equation of ' A '

Characteristic Roots (or) Eigen Values (or) Latent Roots:

Let ' A ' be a square matrix ' A ' is called determinant then the roots of $|A-\lambda I|=0$ are called as characteristic roots (or) Eigen values (or) Latent roots.

Characteristic Vector (or) Eigen Vector:

Let A be a square matrix and λ is a characteristic root. If $\bar{x}$ is a non-zero vector such that

$A\bar{x}=\lambda\bar{x}$ then $\bar{x}$ is called characteristic vector corresponding to characteristic root λ.

Note: $A\bar{x}=\lambda\bar{x}$

$A\bar{x}-\lambda\bar{x}=0 \Rightarrow (A-\mathrm{I}\lambda)\bar{x}=0$

Cayley - Hamilton Theorem:

Statement: Every Square matrix satisfies its characteristic equations.

Proof: Let ' A ' be a square matrices of order ' n ' The characteristic equation of ' A ' is

$|A-\lambda I|=0$

$$\Rightarrow |A-\lambda I|=(-1)^n[\lambda^n+a_1\lambda^{n-1}+a_2\lambda^{n-2}+a_3\lambda^{n-3}+\cdots+ a_{n-1}\lambda+a_n]=0 \rightarrow \quad (1)$$

To prove

$$(-1)^n[A^n+a_1A^{n-1}+a_2A^{n-2}+a_3A^{n-3}+\cdots\cdots+a_{n-1}A+a_n\mathrm{I}]=0 \rightarrow \quad (2)$$

Every element of $(A-\lambda I)$ is a polynomial λ of degrees almost 1.

Every element of $\operatorname{adj}(A-\lambda I)$ is a polynomial in ' λ ' of degrees $(n-1)$ (or) less

$\Rightarrow \operatorname{adj}(A-\lambda I)=B_1\lambda^{n-1}+B_2\lambda^{n-2}+B_3\lambda^{n-3}+\cdots+B_{n-1}\lambda+B_n$

where $B_1, B_2, B_3, \ldots\ldots B_{n-1}, B_n$ are square matrix of order ' n '.

we know that any matrix $'n'$

we have

$$\begin{gathered}A\cdot\operatorname{adj}A=|A|\cdot I.\\ \text{put } A=\mathrm{A}-\lambda I.\\ (A-\lambda I)\operatorname{adj}(A-\lambda I)=|A-\lambda I|\cdot I\ [\because\ \text{from oqn }(1)]\end{gathered}$$

$$\Rightarrow\ (A-\lambda I)[B_1\lambda^{n-1}+B_2\lambda^{n-2}+\cdots\cdots+B_{n-1}\lambda+B_n]=(-1)^n$$
$$[\lambda^n+a_1\lambda^{n-1}+a_2\lambda^{n-2}+\cdots\cdots+a_{n-1}\lambda+a_n]I.$$

compare like powers of λ

$$\begin{aligned}-B_1&=(-1)^nI\ (\lambda^n\text{ coefficients })\\ AB_1-B_2&=(-1)^na_1I\ (\lambda^{n-1}\text{ coetticients })\\ &\vdots\\ AB_{n-1}-B_n&=(-1)^na_{n-1}I\ \ (\lambda\text{-coefficients })\\ AB_n&=(-1)^na_nI\ \ (\text{constant form})\end{aligned}$$

Multiply above $(n+1)$ eqn's with $A^n, A^{n-1}, \ldots\ldots, AI$ respective, we get

$$\begin{aligned}-A^nB_1&=(-1)^nA^n\\ A^nB_1-A^{n-1}B_2&=(-1)^n\cdot a_1A^{n-1}\\ &\vdots\\ A^2B_{n-1}-B_nA&=(-1)^na_{n-1}A\\ A\cdot B_n&=(-1)a_n\mathrm{I}\cdot\end{aligned}$$

adding the above $(n+1)$ equations we get,

$(-1)^n[A^n+a_1A^{n-1}+\cdots+a_{n-1}\cdot A+a_nI]=0$

Every square matrix will be satisfied its characteristics equation.

Hence the theorem is proved.

3.3. QUADRATIC FORMS:

Quadratic Form:

An expression of the form $\sum_{i=1}^n\sum_{j=1}^n a_{ij}x_ix_j$ where a_{ij} and real number is called Quadratic form.

Eg: $2x_1^2+3x_2^2+4x_3^2+3x_1x_2+6x_2x_3+4x_3x_1$ is a quadratic form of three variables x_1, x_2, x_3.

Matrix of Quadratic Form:

Let $\phi = x^{\top} \cdot Ax$ is a real quadratic form where $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. Then ' A ' is called Matrix of the quadratic form.

Note: The matrix of quadratic form ' A ' is always Symmetric form

Reduction of A Real Quadratic Form:

If A be any ' n ' rows real symmetric matrix of rank r then $\exists$ a real non-singular matrix. $p \exists p^T AP = \text{diag}[1,1,1, \dots ,1, -1, \dots, -1,0,0]$

Cochran's Theorem:

Let $x_1, x_2, \dots, x_n$ be the random sample drawn from normal population with parameter

(0, $\theta^{k)}$. Let the sum of the squares of the this values to written in the form. $\sum x_i^2 = \theta_1 + \theta_2 + \cdots \cdots + \theta_k$ where θ_j is the quadratic form $x_1 + x_2, \dots \dots x_n$ with rank $\frac{(r_j)}{j}$

$j = 1,2, \dots k$ then the random voriables $\theta_1, \theta_2, \dots \theta_k$ are mutually independent and θ_j / σ^2 is x^2-variate with r_j degrees of freedom if $\sum_{j=1}^{k} r_j = n$.

1) Prove that value of independent matrix $(A^2 = A)$ are always either zeros (or) ones.

Sol): Let $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

Now

$$A^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = A$$

$$A^2 = A$$

$\therefore A$ is idempotent

Now

$$A - \lambda I = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - \lambda & 0 + \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} - \lambda \end{bmatrix}$$

characteristic equation of ' A ' is

$$|A-\lambda I| = 0$$

$$\begin{bmatrix} \frac{1}{2}-\lambda & 0+\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}-\lambda \end{bmatrix} = 0$$

$$(\frac{1}{2}-\lambda)^2 - \frac{1}{4} = 0$$

$$\frac{1}{4} + \lambda^2 - \lambda - \frac{1}{4} = 0$$

$$\lambda^2 - \lambda = 0$$

$$\lambda(\lambda-1) = 0$$

$\lambda = 0\,;\lambda - 1 = 0 \Rightarrow \lambda = 1$

$\lambda = 0{,}1.$

The Eigen values of independent matrix are 0's (or) 1's

2) Find the characteristic roots are of $A' = \begin{bmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix}$

Sol: Given $\begin{bmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix}$

$$A-\lambda I = \begin{bmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} + \begin{bmatrix} 6-\lambda & -2 & 2 \\ -2 & 3-\lambda & -1 \\ 2 & -1 & 3-\lambda \end{bmatrix}$$

Characteristic εq^n of A is $|A-\lambda I| = 0$

$$\begin{vmatrix} 6-\lambda & -2 & 2 \\ -2 & 3-\lambda & -1 \\ 2 & -1 & 3-\lambda \end{vmatrix} = 0$$

$$6-\lambda[(3-\lambda)^2-(1)] + 2[-2(3-\lambda)+2] + 2[2-2(3-\lambda)] = 0$$

$$6-\lambda[9+\lambda^2-6\lambda-1] + 2[-6+2\lambda+2] + 2[2-6+2\lambda] = 0$$

$$6-\lambda[\lambda^2-6\lambda+8] + 2[2\lambda-4] + 2[2\lambda-4] = 0$$

$$6\lambda^2 - 36\lambda + 48 - \lambda^3 + 6\lambda^2 - 8\lambda + 4\lambda - 8 + 4\lambda - 8 = 0$$

$$-\lambda^3 + 12\lambda^2 - 36\lambda + 32 = 0$$

$$\lambda^3 - 12\lambda^2 + 36\lambda - 32 = 0$$

$\lambda = 2$, satisfies the εq^n

$$(\lambda-2)(\lambda-2)(\lambda-8) = 0$$

$$\lambda = 2{,}2{,}8$$

characteristic vector corresponding to $\lambda = 2$.

Let $\bar{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ be the characteristic vector corresponding to $\lambda = 2$

$$[A - \lambda I]\bar{x} = 0$$
$$(A - 2I)\bar{x} = 0$$

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

It reduces to Normal Form

$$R_2 \longrightarrow 2R_2 + R_1;\ R_3 \longrightarrow 2R_3 - R_1$$
$$= \begin{bmatrix} 4 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$\because (\mathrm{A} - 2I)\bar{x} = 0$$
$$\begin{bmatrix} 4 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0$$
$$\begin{bmatrix} 4x - 2y + 2z \\ 0 \\ 0 \end{bmatrix} = 0$$
$$4x - 2y + 2z = 0$$
$$2x - y + z = 0$$

put $z = k_1$ and $y = k_2$ then

$$2x - k_2 + k_1 = 0$$
$$2x = k_2 - k_1$$
$$x = \frac{k_2 - k_1}{2}$$

$\bar{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{k_2 - k_1}{2} \\ k_2 \\ k_1 \end{bmatrix}$ be the corresponding characteristic vector corresponding to $\lambda = 2$.

characteristic vector corresponding to $\lambda = 8$

$$[A - \lambda I]\bar{x} = 0$$

$$[A-8I]\bar{x}=0$$

$$A-8I=\begin{bmatrix}6&-2&2\\-2&3&-1\\2&-1&3\end{bmatrix}-\begin{bmatrix}8&0&0\\0&8&0\\0&0&8\end{bmatrix}=\begin{bmatrix}-2&-2&2\\-2&-5&-1\\2&-1&-5\end{bmatrix}$$

$R_2\to R_2-R_1; R_3\to R_3+R_1$

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

$R_3\to R_3-R_2$

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

$$[A-8\mathrm{I}]\bar{x}-0\Rightarrow\begin{bmatrix}-2&-2&2\\0&-3&-3\\0&0&0\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}=0$$

$$\begin{bmatrix}-2x-2y+2z\\-3y-3z\\0\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

$$\Rightarrow\ -2x-2y+2z=0$$

$$x+y-z=0$$

put $z=k\Rightarrow y=-k$

put $y=-k, z=k$, in

$x-k-k=0\ \ \Rightarrow x=2k$

$$\bar{x}=\begin{bmatrix}x\\y\\z\end{bmatrix}=\begin{bmatrix}2k\\-k\\k\end{bmatrix}=k\begin{bmatrix}2\\-1\\1\end{bmatrix}$$

$\bar{x}=k\begin{bmatrix}2\\-1\\1\end{bmatrix}$ be the characteristic vector corresponding to $\lambda=8$

Similar Matrices and Equivalent Vectors:

1) Show that $(AB)^{\top}=B^{\top}\cdot A^{\top}$

Sol: Let $A=\left[a_{ij}\right]_{m\times n}; B=\left[b_{jk}\right]_{n\times p}$

Now $AB=\left[a_{ij}\right]_{m\times n}\left[b_{jk}\right]_{n\times p}$

$AB=[c_{ik}]_{m\times p}$ where $c_{ik}=\sum_{j=1}^{n}\ a_{ij}b_{jk}$

Now $(AB)^{\top}=[C_{ki}]$ where $F_{ki}=C_{ik}$.

$A^{\top}=\left[a_{ij}\right]_{m\times m}$ where $a'_{ji}=a_{ij}$.

$B^{\top}=\left[b'_{kj}\right]_{p\times n}$ where $b'_{kj}=b_{ij}$.

Now $B^{\top} \cdot A^{\top} = [b'_{kj}]_{p\times n}[a'_{ij}]_{n\times m}$

$$=[c_{ki}]_{p\times n} \text{ where } c_{ki} = \sum_{j=1}^{n} b'_{kj} a'_{ji}$$

$$Now\ c'_{ki} = c_{ik} \sum_{j=1}^{k} a_{ij} b_{jk}$$

$$= \sum_{j=1}^{n} a'_{ji} b'_{jk} = c_{ki}$$

$$\therefore (AB)^{\top} = B^{\top} A^{\top}$$

2) Show that $(AB)^{-1} = B^{-1} \cdot A^{-1}$.

Sol) Now $(AB)(B^{-1} \cdot A^{-1}) = A(BB^{-1}) \cdot A^{-1}$

$$= AIA^{-1}$$
$$= AA^{-1}$$
$$= \mathrm{I}$$

Now $(B^{-1} \cdot A^{-1})(AB) = B^{-1}(A^{-1}.A)B$

$$= B^{-1} \cdot IB = BB^{-1} = \mathrm{I}$$
$$(AB)(B^{-1} \cdot A^{-1}) = (B^{-1} \cdot A^{-1})(AB) = \mathrm{I}$$
$$(AB)^{-1} = B^{-1} \cdot A^{-1}$$

Hence proved.

3) It ' A ' is square matrix then show that adj $\mathrm{A}^{\top} = (AdjA)^{\top}$

Sol: Let ' A ' be a square matrix order ' n '.

Then adj$A^{\top}$, (Adj$A)^{\top}$ are square matrix of order $'n'$

ij^{th} element of $(\mathrm{adjA})^{\top}$.

$$\Rightarrow \quad P(B) \leq P(AB) + n - P(A)$$
$$\Rightarrow \quad P(A) + P(B) - n \leq P(AB)$$
$$P(AB) = P(A) + P(B) - n \quad \text{(or)}$$
$$\text{Rank}(AB) = \text{Rank}(A) + \text{Rank}(B) - n$$

Hence proved.

3.4. PROBLEMS:

1) Find the matrix of the quadratic form $x_1^2 + 2x_2^2 - 5x_3^2 - x_1x_2 + 4x_2x_3 - 3x_1x_3$.

Sol): Let $\emptyset = x_1^2 + 2x_2^2 - 5x_3^2 - x_1x_2 + 4x_2x_3 - 3x_1x_3$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & -1/2 & -5/2 \\ 1/2 & 2 & 2 \\ -3/2 & 2 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$x^{\mathsf{T}} \cdot Ax$$

where $\begin{bmatrix} 1 & -1/2 & -5/2 \\ -1/2 & 2 & 2 \\ -3/2 & 2 & -5 \end{bmatrix}$ is the matrix of given quadratic form.

2) Find the quadratic form for the matrix

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

Sol): Given $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 3 \\ 3 & 3 & 1 \end{bmatrix}$

Let $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

Quadratic form $\emptyset = X^{\mathsf{T}}AX$.

$$\begin{aligned}
\emptyset &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 3 \\ 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} x_1 + 2x_2 + 3x_3 \\ 2x_1 + 3x_3 \\ 3x_1 + 3x_2 + x_3 \end{bmatrix} \\
&= x_1(x_1 + 2x_2 + 3x_3) + x_2(2x_1 + 3x_3) + x_3(3x_1 + 3x_2 + x_3) \\
&= x_1^2 + 2x_1x_2 + 3x_1x_3 + 2x_1x_2 + 3x_2x_3 + 3x_1x_3 + 3x_2x_3 + x_3^2 \\
&= x_1^2 + x_3^2 + 4x_1x_2 + 6x_1x_3 + 6x_2x_3
\end{aligned}$$

3) Find the rank, signature index transformed form and normal form of given quadratic form.

$$\begin{aligned}
\emptyset = {} & x_1^2 + 6x_1x_2 + 4x_1x_3 + 2x_1x_4 + 8x_1x_5 + 10x_2^2 + 6x_2x_3 + 8x_1x_4 \\
& + 26x_2x_5 + 12x_3^2 + 8x_3x_4 + 20x_3x_5 + 2x_4^2 + 10x_4x_5 + 17x_5^2
\end{aligned}$$

Sol): Given Quadratic form is

$$\emptyset = x_1^2 + 6x_1x_2 + 4x_1x_3 + 2x_1x_4 + 8x_1x_5 + 10x_2^2 + 6x_2x_3 + 8x_2x_4 + 26x_2x_5 + 12x_3^2 + 8x_3x_4 + 20x_3x_5 + 2x_4^2 + 10x_4x_5 + 17x_5^2$$

$$[\,x_1 \quad x_2 \quad x_3 \quad x_4] = \begin{bmatrix} 1 & 3 & 2 & 1 & 4 \\ 3 & 10 & 8 & 4 & 13 \\ 2 & 8 & 12 & 4 & 10 \\ 1 & 4 & 4 & 2 & 5 \\ 4 & 13 & 10 & 5 & 17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

$$\Rightarrow x^{\mathrm{T}}Ax \text{ where } A = \begin{bmatrix} 1 & 3 & 2 & 1 & 4 \\ 3 & 10 & 8 & 4 & 13 \\ 2 & 8 & 12 & 4 & 10 \\ 1 & 4 & 4 & 2 & 5 \\ 4 & 13 & 10 & 5 & 17 \end{bmatrix}$$

Let $A = I^{\mathrm{T}}AT$ (or) I_5AI_5

$$\begin{bmatrix} 1 & 3 & 2 & 1 & 4 \\ 3 & 10 & 8 & 4 & 13 \\ 2 & 8 & 12 & 4 & 10 \\ 1 & 4 & 4 & 2 & 5 \\ 4 & 13 & 10 & 5 & 17 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathrm{A} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$R_2 \to R_2 - 3R_1;\ R_3 \to R_3 - 2R_1;\ R_4 \to R_4 - R_1;\ R_5 \to R_5 - 4R_1$

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

$C_2 \to C_2 - 3C_1;\ C_3 \to C_3 - 2C_1; C_4 \to C_4 - C_1; C_5 \to C_5 - 4C_1$

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

$R_3 \to R_3 - 2R_2;\ \ R_4 \to R_4 - R_2;\ \ R_5 \to R_5 - R_2$

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

$C_3 \to C_3 - 2C_2;\ C_4 \to C_4 - C_2;\ C_5 \to C_5 - C_2$

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

$R_3 \to R_3/2$ and $C_3 \to C_3/2$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 & 0 \\ 2 & -1 & \frac{1}{2} & 0 & 0 \\ 2 & -1 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 \end{bmatrix} \mathrm{A} \begin{bmatrix} 1 & -3 & 2 & 2 & -1 \\ 0 & 1 & -1 & -1 & -1 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= PAQ$$

Rank of $A =$ no. of non-singular sub-square matrix form $= 3$

Index of $A = P =$ number of positive $(+\text{ ve })$ rows in a diagonal form $= 3$

Signature of $A = S = 2p - \gamma$

$= 6 - 3 = 3.$

Normal form of $\emptyset$ is $y_1^2 + y_2^2 + y_3^2 + 0y_1^2 + 0 \cdot y_0^2$ transformed form is

$$X = QY$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 2 & 2 & -1 \\ 0 & 1 & -1 & -1 & -1 \\ 0 & 0 & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} y_1 - 3y_2 + 2y_3 + 2y_4 - y_5 \\ y_2 - y_3 - y_4 - y_5 \\ -y_{\frac{3}{2}} \\ y_4 \\ y_5 \end{bmatrix}$$

$$\begin{aligned} x_1 &= y_1 - 3y_2 + 2y_3 + 2y_4 - y_5 \\ x_2 &= y_2 - y_3 - y_4 - y_5 \\ x_3 &= \frac{-1}{2} y_3 \\ x_4 &= y_4 \\ x_5 &= y_5 \end{aligned}$$

4) If $AB = A, BA = A$ then prove that A, B are Idempotent Matrix.

Sol): Given that $AB = A, BA = B$

since $AB = A$

$$\begin{aligned} \Rightarrow A(BA) &= A \qquad (\because B = BA \\ (AB)A &= A \qquad\quad AB = A) \\ A \cdot A &= A \\ A^2 &= A \end{aligned}$$

'A' is idempotent

$$BA = B$$
$$B(AB) = B$$
$$(BA)\cdot B = B$$
$$B \cdot B = B$$
$$B^2 = B$$

$\therefore$ 'B' is idempotent.

5) Determinant of Matrix:

Let

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

$\det A = 2(24-12) - 3(30-6) - 1(10-4)$

$= 24 - 72 - 6 \neq 0$

Properties:

(i) $A^{\mathsf{T}} = |A|$

(ii) Adj A $= |A|$

(iii) $|AB| = |A| \cdot |B|$

$(iv)\;\; A^{-1} = \frac{Ad_jA}{|A|}$

(v) $|A^{-1}| = \dfrac{1}{|A|}$

(vi) If ' A ' is invertable, where $|A| \neq 0$ (or) A is non-singular matrix.

Adjoint of Matrix:

Let ' A ' be a square matrix then the transpose of the cofactor matrix of ' A ' is called Adjoint matrix of A. It is denoted by Adj A.

i.e., adj$A$ = [co-factor of $A]^{-1}$.

Problems on Characteristic roots (or) Eigen Roots (or) Latent Roots:

1) Show that the characteristic roots of diagonal matrix are same as its diagonal element.

Sol): Let $A = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & \mathrm{c} \end{bmatrix}$ be the diagonal matrix.

Now,

$$\mathrm{A}-\lambda I=\begin{bmatrix}a&0&0\\0&b&0\\0&0&c\end{bmatrix}-\lambda\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$
$$=\begin{bmatrix}a&0&0\\0&b&0\\0&0&c\end{bmatrix}-\begin{bmatrix}\lambda&0&0\\0&\lambda&0\\0&0&\lambda\end{bmatrix}$$
$$A-\lambda I=\begin{bmatrix}a-\lambda&0&0\\0&b-\lambda&0\\0&0&c-\lambda\end{bmatrix}$$

characteristic eq^n of ' A ' is $|A-\lambda I|=0$

$$\begin{vmatrix}a-\lambda&0&0\\0&b-\lambda&0\\0&0&c-\lambda\end{vmatrix}=0$$
$$(a-\lambda)[(b-\lambda)(c-\lambda)-0+0]=0$$
$$(a-\lambda)(b-\lambda)(c-\lambda)=0$$
$$a=\lambda,b=\lambda,c=\lambda$$

$\lambda=a,b,c$ are the characteristic roots of A.

The characteristic roots of a diagonal matrix is same as its diagonal elements.

2) Show that characteristic roots of a triangular matrix are same as its diagonal elements.

Sol: Let $A=\begin{bmatrix}a&b&c\\0&d&e\\0&0&f\end{bmatrix}$ be the upper triangular matrix

Now $A-\lambda I=\begin{bmatrix}a-\lambda I&b&c\\0&d-\lambda&e\\0&0&f-\lambda\end{bmatrix}$

3) If A is idempotent matrix the rank of A = Trace of A

Sol: Given A bean idempotent matrix

i.e., $A^2=A$

Eg: 1

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

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

$$A^2 = A$$

$\therefore$ $'A'$ is idempotent matrix.

Trace of A = 1+1 = 2

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

Rank of A =2; Rank of A = Trace of A

Eg 2:

Let $A = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

Now $A^2 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$ = A

$$A^2 = A$$

A is Idempotent matrix

Trace of $A = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = \frac{3}{3} = 1$

Now $A = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

$$R_1 \to \frac{R_1}{\frac{1}{3}}\ ;\ \ R_2 \to \frac{R_2}{\frac{1}{3}}\ ;\ \ R_3 \to \frac{R_3}{\frac{1}{3}}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$R_2 \to R_2 - R_1;\ \ R_3 \to R_3 - R_1$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$C_2 \to C_2 - C_1;\ \ C_3 \to C_3 - C_1$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \sim \begin{bmatrix} I_1 & 0 \\ 0 & 0 \end{bmatrix}$$

Rank of A = 1

Rank of A = Trace of A

∗. Find rank and Inverse of A and also Cauley Hamilton Theorem.

$$A = \begin{bmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & 7 \end{bmatrix}$$

Sol: Given matrix

$$A = \begin{bmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & -2 \\ 6 & 3 & 0 & 7 \end{bmatrix}$$

$$R_2 \to 2R_2 - R_1 \;\; ; \qquad R_3 \to 2R_3 - 3R_1 \, ; \qquad R_4 \to R_4 - 3R_1$$

$$\begin{bmatrix} 2 & 3 & -1 & -1 \\ 0 & -5 & -3 & -7 \\ 0 & -7 & 9 & -1 \\ 0 & -6 & 3 & 10 \end{bmatrix}$$

$$C_2 \to 2C_2 - 3C_1 \;\; ; \qquad C_3 \to 2C_3 + C_1 \, ; \quad C_4 \to 2C_4 + C_1$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & -6 & -14 \\ 0 & -14 & -18 & -2 \\ 0 & -12 & 6 & 20 \end{bmatrix}$$

$$R_3 \to 10R_3 - 14R_2 \;\; ; \qquad R_4 \to 10R_4 - 12R_2$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & -6 & -14 \\ 0 & 0 & 264 & 176 \\ 0 & 0 & 132 & 368 \end{bmatrix}$$

$$C_3 \to 10C_3 - 6C_2 \;\; ; \qquad C_4 \to 10C_4 - 14C_2$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 2640 & 1760 \\ 0 & 0 & 1320 & 3680 \end{bmatrix}$$

$$C_3 \to C_3 \div 110 \;\; ; \qquad C_4 \to C_4 \div 10$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 24 & 176 \\ 0 & 0 & 12 & 368 \end{bmatrix}$$

$$C_3 \rightarrow C_3 \div 12 \;; \quad C_4 \rightarrow C_4 \div 16$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 2 & 11 \\ 0 & 0 & 1 & 23 \end{bmatrix}$$

$$C_4 \rightarrow 2C_4 - 11C_3$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 33 \end{bmatrix}$$

$$C_4 \rightarrow C_4 \div 33$$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -10 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$C_4 \rightarrow C_4 - C_3 \;; \quad R_1 \rightarrow R_1 \div 2$$
$$R_2 \rightarrow R_2 \div (-10) \;; \quad R_3 \rightarrow R_3 \div 2$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$R_4 \rightarrow R_4 - R_3$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \sim \begin{bmatrix} I_3 & 0 \\ 0 & 0 \end{bmatrix}$$

Rank of matrix A is 3

3.5. CONCLUSION:

The Cayley–Hamilton theorem, quadratic forms, and Cochran's theorem together highlight the power of matrix theory in understanding both algebraic and statistical problems. The Cayley–Hamilton theorem gives a practical method for simplifying matrix computations by allowing a matrix to satisfy its own characteristic equation. Quadratic forms provide a systematic way to study the nature of multivariable expressions and classify them through reduction methods, helping identify whether a system is stable or variable. Cochran's theorem offers a useful technique for breaking down sums of squares into independent components, forming the basis for variance partitioning in statistical models. Together, these methods reinforce the connection between theory and application, strengthening our analytical and problem-solving skills.

3.6. SELF-ASSESSMENT QUESTIONS:

1) Discuss the classification of quadratic forms.
2) Discuss the role of eigenvalues in studying quadratic forms.
3) Explain the importance of the Cayley–Hamilton theorem in linear algebra.
4) Explain how the Cayley–Hamilton theorem is used to find powers of a singular matrix.

3.7. SUGGESTED READINGS:

1) **Hoffman, K. & Kunze, R.** *Linear Algebra*, **Prentice-Hall.**
2) **Strang, G.** *Introduction to Linear Algebra*, **Wellesley-Cambridge Press.**
3) **Lang, S.** *Linear Algebra*, **Springer-Verlag.**
4) **Horn, R.A. & Johnson, C.R.** *Matrix Analysis*, **Cambridge University Press.**
5) **Roman, S.** *Advanced Linear Algebra*, **Springer (Graduate Texts in Mathematics).**
6) **Searle, S.R.**, *Linear Models*, **Wiley** - for Cochran's theorem and quadratic forms in statistics.

Dr. Bala Naga Hima Bindu, Inampudi

LESSON-4

THEORY OF LINEAR ESTIMATION AND LINEAR MODELS

4.0. OBJECTIVES:

After studying this lesson, you should be able to:

- Understand the concept of linear estimation.
- Explain the general linear statistical model.
- Identify assumptions of linear models.
- Define the estimability of linear parametric functions.
- Distinguish between estimable and non-estimable functions.

STRUCTURE

4.1. INTRODUCTION:

In statistics, we often collect sample data to estimate unknown population parameters such as means, regression coefficients, or treatment effects. When the estimator can be written as a **linear combination of the observed data**, the problem belongs to the **theory of linear estimation**.

Linear estimation is important because:

- Many statistical methods (regression, ANOVA, experimental design) can be expressed using linear models.
- Linear estimators are mathematically simple and easy to analyse.
- Under suitable conditions, linear estimators possess optimal properties, including **minimum variance**.

Thus, linear models provide a **unified framework** for statistical inference in practical problems.

Statistical inference is primarily concerned with drawing conclusions about unknown population parameters based on observed sample data. In many practical situations-such as agricultural experiments, industrial quality control, medical research, economics, and social sciences-the relationship between observations and unknown parameters can be expressed in a **linear form**. The study of such problems is known as the **theory of linear estimation**.

In linear estimation, the estimator of an unknown parameter or a function of parameters is assumed to be a **linear function of the observed random variables**. That is, the estimator can be written as a weighted sum of observations. This restriction to linear estimators simplifies mathematical analysis and allows the derivation of important optimality properties, such as minimum variance among a given class of estimators.

Linear estimation plays a central role in statistics because many widely used techniques-such as **regression analysis, analysis of variance (ANOVA), and experimental design-**can all be formulated within the framework of a **general linear statistical model**. By using a common model, diverse statistical methods can be studied in a unified manner.

Another important concept in linear models is **estimability**. In certain models, especially when the design matrix does not have full rank, it may not be possible to estimate all individual parameters uniquely. However, some linear combinations of parameters may still be estimable. Understanding which parametric functions are estimable is essential for meaningful statistical inference.

The theory of linear estimation also provides the foundation for advanced results such as the **Gauss-Markov theorem**, which identifies the best linear unbiased estimator (BLUE), and its generalisation, known as **Aitken's theorem**, applicable when errors are correlated or have unequal variances. Hence, the study of linear estimation is fundamental to both theoretical development and practical application of statistical methods.

4.2. THEORY OF LINEAR ESTIMATION:

Let

$$Y = (Y_1, Y_2, \dots, Y_n)'$$

be a vector of observations.

A **linear estimator** of a parametric function is of the form:

$$\hat{\theta} = a'Y$$

where

$$a = (a_1, a_2, \dots, a_n)'$$

is a vector of known constants.

Properties:

Expectation of a Linear Estimator

$$E(\hat{\theta}) = E(a'Y) = a'E(Y)$$

If

$$E(Y) = \mu$$

then

$$E(\hat{\theta}) = a'\mu$$

Variance of a Linear Estimator

$$Var(\hat{\theta}) = Var(a'Y) = a'Var(Y)a$$

If

$$Var(Y) = \sigma^2 I$$

then

$$Var(\hat{\theta}) = \sigma^2 a'a$$

- Expectation: $E(\hat{\theta}) = a'E(Y)E(\hat{\theta}) = a'E(Y)$
- Variance: $Var(\hat{\theta}) = a'Var(Y)aVar(\hat{\theta}) = a'Var(Y)a$

Example 1 (Linear Estimator)

Let

$$Y_1, Y_2 \sim (\beta, \sigma^2)$$

Consider the estimator:

$$\hat{\beta} = \frac{1}{2}(Y_1 + Y_2)$$

Expectation

$$E(\hat{\beta}) = \frac{1}{2}[E(Y_1) + E(Y_2)] = \frac{1}{2}(\beta + \beta) = \beta$$

Hence, $(\hat{\beta})$ is **unbiased**.

Variance

$$Var(\hat{\beta}) = \frac{1}{4}[Var(Y_1) + Var(Y_2)] = \frac{1}{4}(2\sigma^2) = \frac{\sigma^2}{2}Var(\hat{\beta}) = \frac{1}{4}[Var(Y_1) + Var(Y_2)] = \frac{1}{4}(2\sigma^2) = \frac{\sigma^2}{2}$$

4.3. LINEAR STATISTICAL MODEL:

The general linear model is:

$$Y = X\beta + \varepsilon$$

where

- Y = vector of observations
- X = known design matrix
- β = vector of unknown parameters
- ε = vector of random errors

Expectation and Variance:

$$E(Y) = X\beta$$

$$Var(Y) = Var(\varepsilon) = \sigma^2 I$$

Example 2 (Linear Model):

Let

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \beta = \binom{\beta_1}{\beta_2} X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \beta = \binom{\beta_1}{\beta_2}$$

Then:

$$E(Y) = X\beta = \begin{pmatrix} \beta_1 \\ \beta_1 \\ \beta_1 \end{pmatrix}$$

So, the model depends **only on** β_1.

Problem-1

Let

$$Y = (Y1, Y2, Y3)'$$

be a vector of observations such that

$$E(Y_i) = \theta, \quad Var(Y_i) = \sigma^2, \quad i = 1,2,3$$

and the $Y_i Y_i$'s are uncorrelated.

Consider the linear estimator:

$$\hat{\theta} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

1) Find $E(\hat{\theta})$
2) Find $Var(\hat{\theta})$
3) Comment on unbiasedness

Solution:

Step 1: Write the estimator in matrix form

$$\hat{\theta} = a'Y$$

where

$$a = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)'$$

Step 2: Expectation

$$E(\hat{\theta}) = a'E(Y)$$

Since

$$E(Y) = (\theta, \theta, \theta)'$$

$$E(\hat{\theta}) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\begin{pmatrix}\theta\\ \theta\\ \theta\end{pmatrix} = \frac{1}{3}(3\theta) = \theta$$

Hence, $\hat{\theta}$ is **unbiased**.

Step 3: Variance

Given

$$Var(Y) = \sigma^2 I$$

$$Var(\hat{\theta}) = a'Var(Y)a$$

$$= \sigma^2 a'a$$

$$= \sigma^2\left[\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right]$$

$$= \sigma^2 \left(\frac{3}{9}\right) = \frac{\sigma^2}{3}$$

Final Answer:

$$E(\hat{\theta}) = \theta, \quad Var(\hat{\theta}) = \frac{\sigma^2}{3}$$

Problem 2: Linear Statistical Model

Consider the linear model:

$$Y = X\beta + \varepsilon$$

where

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Assume:

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I$$

1) Find E(Y)
2) Identify the parametric function involved
3) Comment on estimability

Solution:

Step 1: Expectation of Y

$$E(Y) = X\beta$$

$$= \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$= \begin{pmatrix} \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix}$$

Step 2: Parametric Function

The model depends only on the linear parametric function:

$$\theta = \beta_1 + 2\beta_2$$

Step 3: Estimability

Since all expected values are the same and depend on

$$\beta_1 + 2\beta_2$$

This linear function is **estimable**, but β_1 and β_2 **cannot be estimated separately**.

Final Answer

$$E(Y) = \begin{pmatrix} \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix}$$

The model depends only on the estimable function

$$\beta_1 + 2\beta_2$$

4.4. ASSUMPTIONS OF THE LINEAR MODEL:

1) **Zero Mean Errors**

$$E(\varepsilon) = 0$$

2) **Constant Variance**

$$Var(\varepsilon) = \sigma^2 I$$

3) **Uncorrelated Errors**

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

4) **Rank of X**

- Full rank → unique parameter estimates
- Rank deficient → some parameters not estimable

4.5. ESTIMABILITY OF LINEAR PARAMETRIC FUNCTIONS:

A **linear parametric function** is:

$$\psi = l'\beta$$

Definition

A function $l'\beta l'\beta$ is **estimable** if there exists a **linear unbiased estimator** $a'Ya'Y$ such that:

$$E(a'Y) = l'\beta$$

Condition for Estimability (Derivation)

$$E(a'Y) = a'E(Y) = a'X\beta$$

For Unbiasedness:

$$a'X\beta = l'\beta$$

This must hold for all β, hence:

$$l' = a'X$$

Therefore, l must lie in the row space of X.

Advantages of Linear Estimation and Linear Models:

1) **Simplicity of Formulation:**

   Linear estimators are simple weighted sums of observations, making them easy to understand, compute, and interpret.

2) **Unified Framework**:

   Many statistical techniques, such as regression analysis, ANOVA, and experimental design, can be expressed using a single linear model.

3) **Mathematical Tractability:**

   Linear models allow closed-form solutions for estimators, variances, and confidence intervals using matrix algebra.

4) **Optimal Properties**:

   Under standard assumptions, linear estimators possess optimal properties such as minimum variance (Gauss–Markov theorem).

5) **No Need for Normality**:

   The Gauss–Markov theorem does not require normality of errors; only first and second moments are needed.

6) **Ease of Extension:**

   Linear models can be easily extended to generalised models (GLS) to handle correlated or heteroscedastic errors.

7) **Wide Applicability**:

   Linear estimation is applicable in agriculture, economics, medicine, engineering, and social sciences.

Disadvantages of Linear Estimation and Linear Models:

1) **Restriction to Linearity:**

   Only estimators linear in observations are considered; nonlinear estimators may sometimes be more efficient.

2) **Dependence on Model Assumptions:**

Violations of assumptions such as homoscedasticity or uncorrelated errors can lead to inefficient estimates.

3) **Estimability Issues:**

In rank-deficient models, not all parameters are estimable, which can complicate interpretation.

4) **Sensitivity to Outliers:**

Linear estimators, especially least squares estimators, can be highly sensitive to extreme observations.

5) **Limited Flexibility:**

Complex nonlinear relationships cannot be adequately modelled using simple linear models.

6) **Inefficiency Under Heteroscedasticity:**

Ordinary least squares estimators are not efficient when error variances are unequal.

7) **Interpretational Difficulties:**

In models with constraints or aliasing, individual parameter estimates may lack clear meaning.

4.6. CONCLUSION:

The theory of linear estimation and linear models provides a systematic approach for estimating unknown parameters and their linear functions based on observed data. By expressing estimators as linear functions of the observations, the theory offers mathematical simplicity and analytical convenience while ensuring meaningful statistical inference. The general linear model serves as a powerful and unifying framework that encompasses regression analysis, analysis of variance, and experimental design.

A key concept in linear models is estimability, which determines whether a parameter or a linear combination of parameters can be uniquely and unbiasedly estimated. In situations where the design matrix does not have full rank, individual parameters may not be estimable; however, certain linear functions of the parameters may still be estimated reliably. Understanding estimability is therefore essential for the correct interpretation of model parameters.

Under appropriate assumptions on the error structure, linear estimation leads to optimal estimators with minimum variance properties, forming the basis for further theoretical developments such as the Gauss–Markov theorem. Overall, the theory of linear estimation and linear models plays a central role in statistical methodology and forms the foundation for many practical and advanced statistical techniques.

- Linear estimation deals with estimators linear in observations

- Linear models provide a unified framework for regression and ANOVA
- Estimability ensures the uniqueness of the estimation
- Not all parametric functions are estimable in rank-deficient models.

4.7. SELF-ASSESSMENT QUESTIONS:

1) What is meant by a linear estimator?
2) Write the general form of a linear statistical model.
3) State the expectation and variance of a linear estimator.
4) Define a linear parametric function.
5) What is meant by estimability in linear models?
6) Explain the theory of linear estimation with suitable examples.
7) Discuss the assumptions of the general linear statistical model.
8) What are meant by estimable and non-estimable parametric functions? Explain with examples.
9) Explain why rank deficiency of the design matrix affects estimability.
10) Discuss the advantages and disadvantages of linear estimation and linear models.
11) Explain, with a suitable example, a situation where individual parameters are not estimable function of parameters is estimable.

4.8. SUGGESTED READINGS:

1) Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York. Searle, S.R. (1971). *Linear Models*. Wiley, New York.
2) Graybill, F.A. (1976). *Theory and Application of the Linear Model*. Duxbury Press.
3) Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
4) Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.
5) Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.

Prof. V.V. Haragopal

LESSON-5

BEST LINEAR UNBIASED ESTIMATOR AND GAUSS–MARKOV THEOREM

5.0. OBJECTIVES:

After studying this lesson, the student should be able to:

- **Understand the concept of unbiased estimation** in linear regression models.
- **Define and explain the Best Linear Unbiased Estimator (BLUE).**
- **Derive the normal equations** using the least squares method.
- **State and interpret the Gauss–Markov theorem** and its assumptions.
- **Solve numerical problems** to obtain BLUE and estimate linear functions of parameters.

STRUCTURE:

5.1. INTRODUCTION:

In many practical situations, the relationship between a dependent variable and one or more independent variables is modeled using a **linear regression model**. The main objective is to estimate the unknown parameters of the model based on observed data.

Several estimators may be constructed for these parameters. However, an estimator should possess desirable properties such as **linearity, unbiasedness and minimum variance**. Among all estimators that are linear functions of the observations and are unbiased, we seek the one with the **smallest variance**. This leads to the concept of the **Best Linear Unbiased Estimator (BLUE)**.

The method of **least squares** provides such an estimator under certain assumptions on the error terms. The **Gauss-Markov theorem** establishes that the least squares estimator is BLUE, making it fundamental in regression analysis and statistical inference.

Thus, this lesson focuses on the derivation, properties, and applications of BLUE and the Gauss–Markov theorem.

Moreover, the concept of BLUE provides a unifying framework for understanding the efficiency of different estimation methods in linear models. By restricting attention to linear and unbiased estimators, the Gauss–Markov theorem offers a clear criterion for optimality based solely on variance minimization. This result not only simplifies theoretical analysis but also guides practical model building and interpretation. As a consequence, BLUE serves as a cornerstone in statistical modeling, enabling reliable parameter estimation across a wide range of applied disciplines.

5.2. LINEAR UNBIASED ESTIMATORS:

Before identifying the best estimator, it is necessary to understand what is meant by a **linear** and an **unbiased** estimator. In regression analysis, estimators are constructed using observed sample data, and their performance is judged based on properties such as simplicity, unbiasedness, and variability. This section introduces the class of estimators that are linear functions of the observations and whose expected values equal the true parameters.

Consider the general linear statistical model:

$$Y = X\beta + \varepsilon,$$

where Y is the vector of observations, X is the known design matrix, β is the vector of unknown parameters, and ε is the random error vector with

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I.$$

This model provides the framework for defining linear unbiased estimators.

5.3. BEST LINEAR UNBIASED ESTIMATOR (BLUE):

In Section 5.2, we discussed the class of linear unbiased estimators. Since there can be many estimators that satisfy linearity and unbiasedness, it becomes necessary to choose the one that is **most efficient**. Efficiency is measured in terms of **variance**. The estimator with the smallest variance among all linear unbiased estimators is called the **Best Linear Unbiased Estimator (BLUE)**.

Definition:

An estimator $\hat{\beta}\ of \beta$ is said to be the **Best Linear Unbiased Estimator (BLUE)** if:

1) It is a **linear function** of the observations Y, i.e.,

$$\hat{\beta} = AY,$$

2) It is **unbiased**, that is,

$$E(\hat{\beta}) = \beta,$$

3) It has the **minimum variance** among all linear unbiased estimators of β.

Least Squares Estimator:

For the linear model:

$$Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I,$$

the estimator obtained by minimizing the sum of squared errors

$$S = (Y - X\beta)'(Y - X\beta)$$

is called the **least squares estimator**.

Differentiating with respect to β and equating to zero:

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta = 0,$$

we obtain the **normal equations**:

$$X'X\hat{\beta} = X'Y.$$

Solving,

$$\hat{\beta} = (X'X)^{-}X'Y.$$

If $X'X$ is nonsingular,

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

BLUE of β

The least squares estimator $\hat{\beta}$ is linear in Y, unbiased, and (as shown by the Gauss–Markov theorem) has minimum variance among all linear unbiased estimators. Hence,

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y}$$

is the **BLUE** of β.

Variance of BLUE

The variance–covariance matrix of $\hat{\beta}$ is given by:

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

This matrix measures the precision of the parameter estimates.

Numerical Example:

From the simple regression example in Section 5.3, we obtained:

$$\hat{\beta} = \begin{pmatrix} 0.33 \\ 1.50 \end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}.$$

Estimate the linear function:

$$\theta = \beta_0 + \beta_1,$$

and find its variance.

Solution:

Here,

$$l = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Estimator:

$$\widehat{l'\beta} = l'\hat{\beta} = [1\ 1]\begin{pmatrix} 0.33 \\ 1.50 \end{pmatrix} = 1.83.$$

Variance:

$$Var(\widehat{l'\beta}) = \sigma^2 l'(X'X)^{-1}l = \sigma^2[1\ 1]\frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

$$= \sigma^2 \cdot \frac{1}{6}(14 - 6 - 6 + 3) = \frac{5}{6}\sigma^2.$$

$$\boxed{\widehat{l'\beta} = 1.83, \quad Var(\widehat{l'\beta}) = \frac{5}{6}\sigma^2.}$$

5.4. GAUSS-MARKOV THEOREM:

The Gauss–Markov theorem is one of the most important results in linear regression analysis. It provides the theoretical foundation for the use of the least squares method by establishing that, under certain assumptions on the error terms, the least squares estimator is the **Best Linear Unbiased Estimator (BLUE)**. In other words, among all estimators that are linear functions of the observations and unbiased for the parameters, the least squares estimator has the minimum variance.

Statement of the Theorem:

Consider the linear model:

$$Y = X\beta + \varepsilon,$$

where

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I.$$

Then the least squares estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

is the **Best Linear Unbiased Estimator (BLUE)** of β.

Assumptions:

The Gauss–Markov theorem holds under the following assumptions:

1. The model is linear in parameters: $Y = X\beta + \varepsilon$.
2. The errors have zero mean: $E(\varepsilon) = \mathbf{0}$.
3. The errors are uncorrelated and have equal variance: $Var(\varepsilon) = \sigma^2 I$.
4. The matrix X has full column rank.

Meaning of the Theorem:

The theorem states that among all estimators of β that are:

- linear in Y, and
- unbiased,

the least squares estimator $\hat{\beta}$ has the **smallest variance–covariance matrix**. Hence, no other linear unbiased estimator can be more efficient than $\hat{\beta}$.

Key Result:

If $\tilde{\beta}$ is any other linear unbiased estimator of β, then:

$$Var(\tilde{\beta}) - Var(\hat{\beta}) \geq 0,$$

that is, the difference is positive semidefinite. Therefore,

$$Var(\tilde{\beta}) \geq Var(\hat{\beta}).$$

This proves that $\hat{\beta}$ is the **BLUE**.

Importance:

The Gauss-Markov theorem justifies the widespread use of the least squares method in regression analysis. It assures us that, under mild conditions, least squares provide the most precise estimates possible within the class of linear unbiased estimators.

5.5. ESTIMATION OF LINEAR FUNCTIONS:

In many practical situations, interest may not be in estimating the entire parameter vector β, but in estimating certain **linear functions** of the parameters, such as sums, differences, or other combinations. This section explains how such functions can be estimated using the BLUE.

Linear Function of Parameters

A linear function of β is of the form:

$$\theta = l'\beta,$$

where *l* is a known vector of constants.

Examples:

- β_1
- $\beta_1 + \beta_2$
- $\beta_2 - \beta_1$

Estimator of $l'\beta$

If $\hat{\beta}$ is the BLUE of β\betaβ, then the estimator of $l'\beta$ is:

$$\widehat{l'\beta} = l'\hat{\beta}.$$

Unbiasedness

Since $E(\hat{\beta}) = \beta,$

$$E(\widehat{l'\beta}) = E(l'\hat{\beta}) = l'E(\hat{\beta}) = l'\beta.$$

Hence, $l'\hat{\beta}$ is an unbiased estimator of $l'\beta$.

Variance

The variance of $\widehat{l'\beta}$ is given by:

$$Var(\widehat{l'\beta}) = l'Var(\hat{\beta})l = \sigma^2 l'(X'X)^{-1}l.$$

Estimability

A linear function $l'\beta$ is said to be **estimable** if there exists a vector a such that:

$$E(a'Y) = l'\beta.$$

That is, l' lies in the row space of X. Only estimable functions can be estimated unbiasedly.

Numerical Example:

From the simple regression example in Section 5.3, we obtained:

$$\hat{\beta} = \begin{pmatrix} 0.33 \\ 1.50 \end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}.$$

Estimate the linear function:

$$\theta = \beta_0 + \beta_1,$$

and find its variance.

Solution:

Here,

$$l = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Estimator:

$$\widehat{l'\beta} = l'\hat{\beta} = [1\ 1]\begin{pmatrix} 0.33 \\ 1.50 \end{pmatrix} = 1.83.$$

Variance:

$$Var(\widehat{l'\beta}) = \sigma^2 l'(X'X)^{-1}l = \sigma^2[1\ 1]\frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

$$= \sigma^2 \cdot \frac{1}{6}(14 - 6 - 6 + 3) = \frac{5}{6}\sigma^2.$$

$$\boxed{\widehat{l'\beta} = 1.83, \quad Var(\widehat{l'\beta}) = \frac{5}{6}\sigma^2.}$$

5.6. ADVANTAGES AND DISADVANTAGES:

Advantages of BLUE

1) **Minimum Variance**
   Among all linear and unbiased estimators, BLUE has the smallest variance, making it the most efficient estimator in this class.

2) **Unbiased Estimation**
   The expected value of the BLUE equals the true parameter value, ensuring no systematic over- or under-estimation.

3) **No Distributional Assumption**
The Gauss–Markov theorem does not require normality of the error terms; only mean zero and constant variance are needed.

4) **Theoretical Foundation of Least Squares**
BLUE provides a strong theoretical justification for using the least squares method in regression analysis.

5) **Wide Applicability**
It is extensively used in economics, engineering, agriculture, biostatistics, and social sciences for reliable parameter estimation.

Disadvantages of BLUE:

1) **Restricted to Linear Estimators**
BLUE is optimal only within the class of linear unbiased estimators; non-linear estimators may sometimes perform better.

2) **Dependence on Model Assumptions**
If assumptions such as homoscedasticity or uncorrelated errors are violated, BLUE may lose its optimality.

3) **Sensitivity to Multicollinearity**
When the design matrix X is nearly singular, BLUE can have large variances and unstable estimates.

4) **Not Necessarily Best Overall Estimator**
If errors are normally distributed, estimators like the Maximum Likelihood Estimator (MLE) may be more efficient than BLUE.

5.7. CONCLUSION:

In this unit, we studied the problem of estimating parameters in a linear regression model using linear unbiased estimators. The main points covered are summarized below:

- The linear model is given by $Y = X\beta + \varepsilon$, with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 I$.

- A **linear estimator** is a linear function of the observations, and it is **unbiased** if its expected value equals the true parameter.

- The **Best Linear Unbiased Estimator (BLUE)** is the linear unbiased estimator with the minimum variance.

- The least squares estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is the BLUE of β .

- The **normal equations** $X'X\hat{\beta} = X'Y$ are obtained by minimizing the sum of squared errors.

- The **Gauss–Markov theorem** proves that the least squares estimator is BLUE under mild assumptions and does not require normality of errors.

- Linear functions of parameters $l'\beta$ can be estimated by $l'\hat{\beta}$, and their variance is $\sigma^2 l'(X'X)^{-1}l$.
- Numerical examples illustrate the computation and application of BLUE in regression problems.

This unit establishes the theoretical foundation for regression analysis and provides tools for efficient estimation and inference in practical applications.

5.8. SELF-ASSESSMENT QUESTIONS:

A. Short Answer Questions:

1) What is meant by an unbiased estimator?
2) Define a linear estimator.
3) What is meant by BLUE?
4) Write the general linear model.
5) What are the normal equations?
6) State the Gauss–Markov theorem.
7) What does "best" mean in BLUE?
8) Does the Gauss–Markov theorem assume normality of errors?
9) Write the expression for the variance of $\hat{\beta}$.
10) What is meant by an estimable function?

B. Descriptive / Long Answer Questions:

1) Explain the concept of linear unbiased estimators.
2) Derive the normal equations using the least squares method.
3) Define BLUE and discuss its properties.
4) State and explain the Gauss–Markov theorem with assumptions.
5) Show that the least squares estimator is unbiased.
6) Explain how linear functions l′βl'\betal′β are estimated and find their variance.
7) Discuss the importance of the Gauss–Markov theorem in regression analysis.

5.9. SUGGESTED READINGS:

The following books and references are recommended for further study and deeper understanding of linear models, BLUE, and the Gauss–Markov theorem:

1) **Rao, C. R.** – *Linear Statistical Inference and Its Applications*, Wiley.

→ A classic reference on estimation theory and linear models.

2) **Montgomery, D. C., Peck, E. A., and Vining, G. G.** – *Introduction to Linear Regression Analysis*, Wiley.

→ Excellent for regression methods and applications.

3) **Draper, N. R. and Smith, H.** – *Applied Regression Analysis*, Wiley.

→ Focuses on practical aspects of regression analysis.

4) **Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W.** – *Applied Linear Statistical Models*, McGraw-Hill.

→ Widely used textbook with examples and exercises.

5) **Seber, G. A. F. and Lee, A. J.** – *Linear Regression Analysis*, Wiley.

→ Covers theory and computation in detail.

6) **Graybill, F. A.** – *Introduction to Matrices with Applications in Statistics*, Wadsworth.

→ Useful for matrix methods used in BLUE.

Prof. V.V. Haragopal

LESSON-6

GENERALIZED LINEAR MODEL AND GENERALIZED GAUSS-MARKOV (AITKEN'S THEOREM)

6.0. OBJECTIVES:

After studying this lesson, you should be able to:

- Understand models with correlated and heteroscedastic errors.
- Explain the concept of Generalized Least Squares (GLS).
- State and interpret Aitken's theorem.
- Compare Ordinary Least Squares (OLS) and GLS estimators.

STRUCTURE:

6.1. INTRODUCTION:

In many practical data analysis problems, the assumptions of the classical linear regression model are often violated. The traditional model assumes that the error terms are independent and identically distributed with zero mean and constant variance. However, in real-world applications such as time-series analysis, econometrics, environmental studies, and engineering experiments, errors may be correlated across observations or may exhibit unequal variances, a situation known as heteroscedasticity. Under such circumstances, the Ordinary Least Squares (OLS) estimator, although still unbiased, no longer possesses the property of minimum variance among all linear unbiased estimators.

To address these limitations, the linear model is extended by allowing a more general form for the variance–covariance matrix of the error vector. This leads to the formulation of the **Generalized Linear Model**, in which the error variance is no longer restricted to a scalar multiple of the identity matrix. The method of **Generalized Least Squares (GLS)** naturally arises from this framework, providing a way to incorporate the known error structure into the estimation process. The theoretical justification for GLS is given by the **Generalized Gauss-**

Markov theorem, also called **Aitken's theorem**, which establishes GLS as the best linear unbiased estimator under generalized error conditions. This lesson focuses on understanding these extensions and their importance in obtaining efficient and reliable parameter estimates in practical statistical modeling.

6.2. GENERALIZED LINEAR MODEL:

The generalized linear model provides a natural extension of the classical linear regression model by relaxing the restrictive assumptions on the error structure. In many practical situations, the variability in observations is not uniform and the errors may exhibit correlation due to time, space, or grouping effects. To capture such realistic features of data, the generalized linear model allows the error term to have a general variance–covariance matrix rather than assuming equal and independent variances.

The model is written as:

$$Y = X\beta + \varepsilon,$$

where

- Y is an $n \times 1$ vector of observed responses,
- X is an $n \times p$ known design matrix,
- $\beta \; is \; a \; p \times 1$ vector of unknown parameters, and
- $\varepsilon \; is \; an \; n \times 1$ vector of random errors.

The key assumptions on the error term are:

$$E(\varepsilon) = 0, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 V,$$

where V is a known n×n **positive definite matrix**.

The matrix V represents the pattern of variances and covariances among the errors. If V=I, the errors are uncorrelated and have equal variances, and the model reduces to the classical linear model. If V is diagonal with unequal elements, the model accounts for heteroscedasticity. If V has non-zero off-diagonal elements, it represents correlated errors.

Thus, the generalized linear model provides a flexible framework for modeling data with non-spherical error structures and forms the basis for deriving efficient estimation procedures such as Generalized Least Squares.

6.3. GENERALIZED LEAST SQUARES (GLS):

In the generalized linear model, the presence of correlated errors or unequal variances makes the Ordinary Least Squares (OLS) method inadequate from the point of view of efficiency. Although OLS estimators remain unbiased under such conditions, they no longer have minimum variance among all linear unbiased estimators. To overcome this limitation, the method of **Generalized Least Squares (GLS)** is employed. GLS modifies the least squares criterion by explicitly incorporating the known variance–covariance structure of the errors, thereby assigning appropriate weights to observations and leading to more precise parameter estimates.

When the error vector satisfies

$$\mathrm{Var}(\varepsilon) = \sigma^2 V,$$

the GLS estimator of β\betaβ is obtained by minimizing the weighted sum of squared residuals:

$$(Y - X\beta)' V^{-1} (Y - X\beta).$$

This yields the estimator:

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1} X'V^{-1}Y.$$

The variance–covariance matrix of the GLS estimator is:

$$\mathrm{Var}(\hat{\beta}_{GLS}) = \sigma^2 (X'V^{-1}X)^{-1}.$$

Thus, GLS takes into account the structure of the error covariance matrix and produces more efficient estimates than OLS whenever errors are correlated or heteroscedastic. An important property of GLS is that when V=I, the GLS estimator reduces to the ordinary least squares estimator, showing that OLS is a special case of GLS.

6.4. NUMERICAL EXAMPLE: GLS

Consider the generalized linear model

$$Y = X\beta + \varepsilon,$$

where

$$X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, Y = \begin{pmatrix} 4 \\ 6 \end{pmatrix},$$

and the variance–covariance matrix of errors is

$$\mathrm{Var}(\varepsilon) = \sigma^2 V, \quad V = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$

Find the **GLS estimate** of β.

Solution:

The GLS estimator is given by:

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1} X'V^{-1}Y.$$

Step 1: Find V^{-1}

$$V^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

Step 2: Compute $X'V^{-1}X$

$$X'V^{-1}X = (1 \ \ 1) \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1 + 0.25 = 1.25.$$

Step 3: Compute $X'V^{-1}Y$

$$X'V^{-1}Y = (1 \quad 1)\begin{pmatrix}1 & 0\\0 & 0.25\end{pmatrix}\begin{pmatrix}4\\6\end{pmatrix} = 4 + 1.5 = 5.5.$$

Step 4: Compute GLS estimate

$$\hat{\beta}_{GLS} = \frac{5.5}{1.25} = 4.4.$$

Answer

$$\boxed{\hat{\beta}_{GLS} = 4.4}$$

Generalized Linear Model (GLM):

Advantages:

1) **Handles non-normal data**

   GLMs allow response variables to follow distributions like binomial, Poisson, or gamma, not only normal.

2) **Flexible relationship**

   The link function connects the mean of the response to predictors, allowing non-linear relationships.

3) **Widely applicable**

   Used in logistic regression, Poisson regression, survival analysis, epidemiology, and social sciences.

4) **Interpretable parameters**

   Coefficients often have meaningful interpretations (e.g., odds ratios in logistic regression).

5) **Unifies many models**

   Linear regression, logistic regression, and Poisson regression are all special cases of GLM.

Disadvantages:

1) **Model selection is difficult**

   Choosing the correct distribution and link function requires experience.

2) **Computationally intensive**

   Estimation is done using iterative methods, which can be slow for large datasets.

3) **Sensitive to misspecification**

   Incorrect choice of link or distribution leads to biased results.

4) **Assumes independence**

   Standard GLMs assume observations are independent, which may not always be true.

5) **Less intuitive for beginners**

   Concepts like link functions and likelihood estimation are harder to understand than simple linear regression.

6.5. GENERALIZED GAUSS-MARKOV THEOREM (AITKEN'S THEOREM):

The efficiency and optimality of the Generalized Least Squares estimator are formally established by the **Generalized Gauss–Markov theorem**, commonly referred to as **Aitken's theorem**. Just as the classical Gauss–Markov theorem shows that the Ordinary Least Squares estimator is the best linear unbiased estimator under the assumption of independent and homoscedastic errors, Aitken's theorem extends this important result to situations where the error terms have a general variance–covariance structure. It provides the theoretical foundation for preferring GLS over OLS in models with correlated or unequal error variances.

Statement:

Consider the **general linear model**

$$Y = X\beta + \varepsilon$$

where

$$E(\varepsilon) = 0, Var(\boldsymbol{\varepsilon}) = \sigma^2 V$$

$with\ Va$ **known positive definite matrix**.

Theorem (Aitken)

Among all **linear unbiased estimators** of β, the estimator

$$\boxed{\hat{\beta}_{GLS} = (\boldsymbol{X'V^{-1}}X)^{-1}\boldsymbol{X'V^{-1}}Y}$$

has **minimum variance**.

Hence, $\hat{\beta}_{GLS}$ is the **Best Linear Unbiased Estimator (BLUE)** of β.

Proof of Aitken's Theorem:

Step 1: Consider a general linear estimator

Let

$$\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$$

where A is a p×n matrix of constants.

Step 2: Condition for unbiasedness

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}E(\boldsymbol{Y}) = \boldsymbol{AX\beta}$$

For unbiasedness:

$$\boldsymbol{AX} = \boldsymbol{I_p} \quad (1)$$

Step 3: Variance of a linear estimator

$$Var(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}Var(\boldsymbol{Y})\boldsymbol{A'} = \sigma^2\boldsymbol{AVA'}$$

Step 4: GLS estimator satisfies unbiasedness

Define

$$\boldsymbol{A_0} = (\boldsymbol{X'V^{-1}}X)^{-1}\boldsymbol{X'V^{-1}}$$

Then:

$$\boldsymbol{A_0X} = (\boldsymbol{X'V^{-1}}X)^{-1}\boldsymbol{X'V^{-1}}X = \boldsymbol{I_p}$$

So, $\hat{\beta}_{GLS} = \boldsymbol{A_0Y}$ is **unbiased**.

Step 5: Compare variances of estimators

Let $\widehat{\boldsymbol{\beta}} = \boldsymbol{AY}$ be any linear unbiased estimator.
Define:

$$\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{A_0}$$

Then:

$$DX=0$$

Now consider:

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\boldsymbol{A_0} + \boldsymbol{D})\boldsymbol{V}(\boldsymbol{A_0} + \boldsymbol{D})'$$

Expanding:

$$= \sigma^2[\boldsymbol{A_0VA_0'} + \boldsymbol{DVD'} + \boldsymbol{A_0VD'} + \boldsymbol{DVA_0'}]$$

Step 6: Cross terms vanish

Since DX=0:

$$\boldsymbol{A_0VD'} = (\boldsymbol{X'V^{-1}}X)^{-1}\boldsymbol{X'D'} = 0$$

Similarly:

$$\boldsymbol{DVA_0'} = 0$$

Thus:

$$Va\,r\left(\hat{\beta}\right) = Var\left(\hat{\beta}_{GLS}\right) + \sigma^2\boldsymbol{DVD'}$$

Step 7: Conclude optimality

Since V is positive definite:

$\sigma^2\boldsymbol{DVD'}$ is positive semi-definite

Therefore:

$$Va\,r\left(\hat{\beta}\right) \geq Var\left(\hat{\beta}_{GLS}\right)$$

Hence, **no other linear unbiased estimator has smaller variance** than the GLS estimator.

This means that among all estimators that are linear functions of the observations and unbiased for β, the GLS estimator has the minimum variance. Hence, Aitken's theorem generalizes the classical Gauss–Markov theorem by replacing the restrictive assumption $\mathrm{Var}(\varepsilon) = \sigma^2 I$ with the more general condition $\mathrm{Var}(\varepsilon) = \sigma^2 V$. When V=I, Aitken's theorem.

Generalized Gauss-Markov Theorem (Aitken's Theorem)

Advantages:

1) **Best Linear Unbiased Estimator (BLUE)**
Aitken's theorem provides the most efficient linear unbiased estimator when error covariance is known.
2) **Handles heteroscedasticity**
Works when error variances are unequal.
3) **Allows correlated errors**
Useful in time-series and spatial data.
4) **Improves efficiency**
Generalized Least Squares (GLS) estimators have smaller variance than OLS.
5) **Extends classical Gauss–Markov theorem**
Makes linear estimation more realistic for practical data.

Disadvantages:

1) **Requires known covariance matrix**
In practice, the error covariance matrix is often unknown.
2) **Estimation becomes complex**
Computing GLS estimators involves matrix inversion and numerical methods.
3) **Sensitive to covariance misspecification**
Incorrect covariance structure leads to inefficient estimates.
4) **Limited to linear models**
Does not handle non-linear mean structures or non-normal responses.
5) **Interpretation remains linear**
Cannot model non-linear relationships between predictors and response.

6.6. COMPARISON OF OLS AND GLS:

Having developed the estimators under both the classical and generalized linear models, it is important to compare the Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) methods. While both aim to estimate the same parameter vector β\betaβ, their performance differs significantly depending on the nature of the error structure. This section highlights the key differences between OLS and GLS and clarifies when GLS should be preferred over OLS in practical applications.

Feature	OLS	GLS
Error variance	$\sigma^2 I$	$\sigma^2 V$
Error structure	Independent and equal variance	Correlated and/or unequal variance
Estimator	$(X'X)^{-1}X'Y$	$(X'V^{-1}X)^{-1}X'V^{-1}Y$
Unbiasedness	Unbiased if $E(\varepsilon) = 0$	Unbiased if $E(\varepsilon) = 0$
Efficiency	BLUE only when V=I	BLUE for general V
Special case	–	Reduces to OLS when V=I
Computation	Simple	More computational effort
Applicability	Classical regression problems	Time-series, panel, spatial, heteroscedastic data

Remarks:

- When errors are independent with equal variance, OLS is optimal and simpler to use.
- When errors are correlated or heteroscedastic, OLS loses efficiency, while GLS remains optimal.
- GLS gives more weight to observations with smaller variances and adjusts for correlations.
- In practice, if V is unknown, it is estimated, leading to Feasible GLS (FGLS).

Thus, the choice between OLS and GLS depends on how well the classical assumptions about the error term are satisfied in a given problem.

6.7. CONCLUSION:

In this lesson, we extended the classical linear regression framework to situations where the usual assumptions about the error structure do not hold. The generalized linear model allows the error terms to be correlated and to have unequal variances, which is often the case in practical data analysis problems. Under such conditions, the Ordinary Least Squares method, although unbiased, is no longer efficient.

To address this issue, the method of **Generalized Least Squares (GLS)** was introduced. By incorporating the known variance–covariance matrix of the errors into the estimation procedure, GLS provides parameter estimates with smaller variance than OLS whenever the errors are heteroscedastic or correlated. The derivation of the GLS estimator shows that it can be obtained by transforming the generalized model into a classical one and then applying OLS.

The optimality of GLS is guaranteed by the **Generalized Gauss–Markov theorem (Aitken's theorem)**, which states that the GLS estimator is the Best Linear Unbiased Estimator of the parameter vector under the generalized model. Finally, a comparison of OLS and GLS highlights that while OLS is simple and effective under classical assumptions, GLS is more appropriate and efficient in realistic situations where those assumptions are violated. Thus, GLS plays a crucial role in modern regression analysis and statistical modeling.

6.8. SELF-ASSESSMENT QUESTIONS:

1) What is meant by a generalized linear model? How does it differ from the classical linear model?
2) Why does the Ordinary Least Squares (OLS) estimator lose efficiency when errors are heteroscedastic or correlated?
3) Derive the Generalized Least Squares (GLS) estimator starting from the weighted least squares criterion.
4) Write down the GLS estimator and its variance–covariance matrix.
5) State the Generalized Gauss-Markov theorem (Aitken's theorem). What is its significance?
6) Explain the meaning of the term **BLUE** in the context of Aitken's theorem.
7) Compare OLS and GLS estimators with respect to assumptions, efficiency, and applicability.
8) In what situations is GLS preferred over OLS? Give practical examples.
9) What is Feasible GLS (FGLS)? Why is it used in practice?
10) Show that GLS reduces to OLS when the variance–covariance matrix of errors is σ2I\sigma^2 Iσ2I.

6.9. SUGGESTED READINGS:

To gain deeper insight into generalized linear models, GLS estimation, and Aitken's theorem, students are encouraged to consult the following standard textbooks and references:

1) **Rao, C.R.** (1973). *Linear Statistical Inference and Its Applications*. Wiley. – A classic reference covering linear models, estimation theory, and extensions of the Gauss–Markov theorem.
2) **Searle, S.R.** (1971). *Linear Models*. Wiley.

   – Provides a comprehensive treatment of linear and generalized linear models with matrix methods.
3) **Graybill, F.A.** (1976). *Theory and Application of the Linear Model*. Duxbury Press.
   – Focuses on both theoretical foundations and practical applications.

4) **Draper, N.R., & Smith, H.** (1998). *Applied Regression Analysis*. Wiley. – An applied perspective on regression, including handling of non-constant variance.

5) **Greene, W.H.** (2018). *Econometric Analysis*. Pearson.

– Extensive coverage of GLS, FGLS, and applications in econometrics.

6) **Montgomery, D.C., Peck, E.A., & Vining, G.G.** (2012). *Introduction to Linear Regression Analysis*. Wiley.

– Useful for understanding practical regression issues and remedies for assumption violations.

Prof. V.V. Haragopal

LESSON-7

ANALYSIS OF VARIANCE (ANOVA)

7.0. OBJECTIVES:

After studying this unit, you should be able to:

- Understand the concept of total variation in ANOVA.
- Explain the decomposition of the total sum of squares.
- Distinguish between one-way and two-way ANOVA.
- Calculate sum of squares due to treatment, error, and interaction effects.
- Analyse both balanced and unbalanced designs in ANOVA.

STRUCTURE:

7.1. INTRODUCTION:

The analysis of variance is a powerful statistical tool for tests of significance. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population, i.e., they have the same mean. For example, five fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is given. We may be interested in finding out whether the effect of these fertilizers in the yields is significantly different or in other words, whether the samples have come for the same normal population. The answer to this problem is provided by the technique of analysis of variance. The basic purpose of the analysis of variance is to test the homogeneity of several means.

The term 'Analysis of Variance' was introduced by Prof. R.A. Fisher in 1920's to deal with problem in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as: (i) Assignable causes, and (ii) Chance causes.

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

Definition:

According to Prof. R.A. Fisher, Analysis of Variance (ANOVA) is the "Separation of variance ascribable to one group of causes from the variance ascribable to other group". By this technique the total variation in the sample data is expressed as the sum of its non-negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause. The ANOVA consists in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to assignable factors (causes) with the estimate due to chance factor (causes), the latter being known as experimental error or simply error.

7.2. ASSUMPTIONS FOR ANOVA TEST:

ANOVA test is based on the test statistics F (or) Variance Ratio. For the validity of the F-test in ANOVA, the following assumptions are made.

i) The observations are independent,

ii) Parent population from which observations are taken is normal, and

iii) Various treatment and environmental effects are additive in nature. In the following sequences we will discuss the analysis of variance for F test

a) One-way classification

b) Two-way classifications

Remarks:

1) ANOVA technique enables us to compare several populations means simultaneously and thus results in lot of savings in terms of time and money as compared to several experiments required for comparing two populations means at a time.

2) As pointed out earlier, the origin of the ANOVA technique lies in agricultural experiments and as such its language is loaded with such terms as treatments, blocks, plots etc. However, ANOVA technique is so versatile that it finds applications in almost all types of design of experiments in various diverse fields such as industry, education, psychology, business etc.

7.3. ANOVA ONE WAY CLASSIFICATION WITH ONE OBSERVATION FOR EACH SUBCLASS:

Layout of One-Way Classification:

A One-Way ANOVA (Analysis of Variance) is used to test whether there are statistically significant differences between the means of three or more independent (unrelated) groups.

$Class\ i$	$Values\ Y_{ij}$	$Total\ T_i$	$Mean\ \bar{Y}_{i.}$
1	$Y_{11}, Y_{12}, \ \dots, Y_{1n_1}$	T_1	$\bar{Y}_{1.}$
2	$Y_{21}, Y_{22}, \ \dots, Y_{2n_2}$	T_2	$\bar{Y}_{2.}$
...	...	...	...
i	$Y_{i1}, Y_{i2}, \ \dots, Y_{in_i}$	T_i	$\bar{Y}_{i.}$
...	...	...	...
k	$Y_{k1}, Y_{k2}, \ \dots, Y_{kn_k}$	T_k	$\bar{Y}_{k.}$
		$Grand\ Total\ G$	$Grand\ Mean\ \bar{Y}_{..}$

Mathematical Model

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Where:

- Y_{ij} is j^{th} observation in the i^{th} sub classes
- $\mu = \ General\ Mean\ Effect$
- $\alpha_i = \ i^{th}\ sub\ class\ additive\ effect$
- $\epsilon_{ij} \sim N(0, \sigma^2)$

Working Rule of ANOVA one-way classification

Explanation of Terms:

- k: number of groups
- N: total number of observations
- Set the Hypothesis
- Degree of Freedom
- SS (Sum of Squares): a measure of variability
  - SS_T = total sum of squares
  - SS_B = sum of squares between groups (explained variation)
  - SS_W = sum of squares within groups (unexplained variation)
  - MS (Mean Square): an average of the sum of squares (SS/df)

- F-ratio: used to determine statistical significance (if F is large enough, the null hypothesis is rejected)

a) Decomposition of Sum of Squares:

Step 1: Set Hypotheses

$NullHypothesis(H_0): All\ group\ means\ \ are\ equal$

$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$

Alternative Hypothesis (H_1): At least one group mean differs

Step 2: Compute Group Totals and Means

- $Yij : Observation\ j\ in\ group\ i$
- $Totalforgroup\ i : T_i = \sum_{j=1}^{n_i} Y_{ij}$
- $Meanforgroup\ i : \overline{Y}_{i}. = \frac{T_i}{n_i}$
- $Grand\ total : G = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$
- $Grand\ mean : \overline{Y} = \frac{G}{N}\ where\ N = \sum_{i=1}^{k} n_i$

Here unknown parameters are $\mu\ and\ t_i$. We estimate the parameters by using the principle of least squares (method of least squares).

Minimize the error sum of squares partially differentiating w.r.t. the parameters.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{----------------(1)}$$

$$\epsilon_{ij} = Y_{ij} - \mu - \alpha_i$$

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {\epsilon_{ij}}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \mu - \alpha_i\right)^2 ------(2)$$

Partial differentiation equ (2) w.r.t μ and equated to zero, we get

$$\frac{\partial E}{\partial \mu} = 0 \rightarrow 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu - \alpha_i\right)(-1) = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu - \alpha_i\right) = 0$$

$$Y_{..} = kn_i\,\mu + n_i\sum_{i=1}^{k}\alpha_i \left(\because \sum \alpha_i = 0\right)$$

$$\mu = \frac{\sum\sum Y_{ij}}{N} = \overline{Y_{..}} \qquad ---------- (3)$$

Partial differentiation equ (2) w.r.t t_i and equated to zero, we get

$$\frac{\partial E}{\partial \alpha_i} = 0 \rightarrow 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu - \alpha_i\right)(-1) = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu - \mathrm{t}_i\right) = 0$$

$$\alpha_{i.} = \overline{Y_{i.}} - \mu$$

$$\alpha_{i.} = \overline{Y_{i.}} - \overline{Y_{..}}$$

$$\epsilon_{ij} = Y_{ij} - \mu - \alpha_i$$

$$\epsilon_{ij} = Y_{ij} - \overline{Y_{i.}}$$

Step 3: Calculate Sum of Squares

Total Sum of Squares (SST):

$$SST_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y_{..}}\right)^2$$

Between Groups Sum of Squares (SSB):

$$SSB_B = \sum_{i=1}^{k} n_i\left(\overline{Y_{i.}} - \overline{Y_{..}}\right)^2$$

Within Groups Sum of Squares (SSW)(Error):

$$SSW_W = \sum\left(Y_{ij} - \overline{Y_{i.}}\right)^2$$

$$SSW_W = SST_T - SSB_B$$

Step 4: Compute Degrees of Freedom

$$df_{between} = k - 1$$

$$df_{within} = N - k$$

$$df_{total} = N - 1$$

Step 5: Compute Mean Squares

$$MSB_B = \frac{SSB}{k - 1}$$

$$MSW_W = \frac{SSW}{N - k}$$

Step 6: Compute F-Ratio

$$F = \frac{MSB_B}{MSW_W}$$

Step 7: Compare with F-Critical

$$Find\ F_{critical}\ from\ F - distribution\ table for given\ \alpha(usually\ 0.05), with$$
$$df_1 = k - 1,\ df_2 = N - k$$

Decision Rule:

$If\ F > F_{critical}: Reject H_0,\qquad If\ F \leq F_{critical}: Fail\ to\ reject H_0$

Step 8: Conclusion

If H_0 is rejected, at least one group mean is significantly different.

One-Way ANOVA Table

$Source\ of\ Variation$	$Sumof\ Squares\ (SS)$	$Degreesof\ Freedom\ (df)$	$MeanSquare\ (MS) = SS/df$	$F - ratio = MS(Between)/MS(Within)$
$Between\ Groups$	SSB_B	$k - 1$	$MSB_B = SSB_B/(k-1)$	$F = MSB_B/MSW_W$
$Within\ Groups$	SSW_W	$N - k$	$MSW_W = SSW_W/(N-k)$	
$Total$	SST_T	$N - 1$		

b) Example for ANOVA One Way:

A researcher wants to test whether three different fertilizers affect plant growth differently. He applies Fertilizer A, B, and C to three groups of plants and records the growth (in cm) after a fixed time.

Data:

Fertilizer	**Plant 1**	**Plant 2**	**Plant 3**
A	20	22	23
B	25	27	26
C	22	20	21

Steps in One-Way ANOVA:

Step 1: Calculate Group Means

- k=3 (number of groups)
- n=3 (observations per group)
- N=9 (total number of observations)

G=206 (Grand Total)

- $\overline{Y_A} = \frac{20+22+23}{3} = 21.67$
- $\overline{Y_B} = \frac{25+27+26}{3} = 26.00$
- $\overline{Y_C} = \frac{22+20+21}{3} = 21.00$

Step 2: Calculate the Overall Mean (Grand Mean)

- $\overline{Y_{..}} = \frac{20+22+23+25+27+26+22+20+21}{9} = \frac{206}{9} = 22.89$

Step 3: Compute Sum of Squares

Total Sum of Squares (SST):

$$SST_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y_{..}}\right)^2 = 52.89$$

$= (20 - 22.89)^2 + (22 - 22.89)^2 + \cdots + (21 - 22.89)^2$

$= 8.35 + 0.79 + 0.01 + 4.45 + 16.87 + 9.67 + 0.79 + 8.35 + 3.57 = 52.857$

(Compute each observation’s squared deviation from the grand mean.)

Between Groups Sum of Squares (SSB):

$$SSB_B = \sum_{i=1}^{k} n_i\left(\overline{Y_{i.}} - \overline{Y_{..}}\right)^2 = 44.19$$

- $FertilizerA: 3(21.67 - 22.89)^2 = 3(1.48) = 4.19$
- $FertilizerB: 3(26 - 22.89)^2 = 3(9.67) = 29.01$
- $FertilizerC: 3(21 - 22.89)^2 = 3(3.57) = 10.71$

$SSB_B = 4.47 + 29.01 + 10.71 = 44.19$

Within Groups Sum of Squares (SSW)(Error):

$$SSW_W = \sum (Y_{ij} - \overline{Y_{i.}})^2 = 8.68$$

$$SSW_W = SST_T - SSB_B = 52.87 - 44.19 = 8.68$$

Step 4: Calculate Degrees of Freedom

- $df_{between} = k - 1 = 3 - 1 = 2$
- $df_{within} = N - k = 9 - 3 = 6$
- $df_{Total} = N - 1 = 9 - 1 = 8$

Step 5: Compute Mean Squares

- $MSB = \frac{SSB}{df_{\text{between}}} = \frac{44.19}{2} = 22.095$
- $MSW = \frac{SSW}{df_{\text{within}}} = \frac{8.68}{6} = 1.447$

Step 6: Calculate the F-Ratio

$$F = \frac{MSB}{MSW} = \frac{22.095}{1.447} \approx 15.24$$

Step 7: Compare with Critical F-Value / Find p-value

- Use F-distribution table or software with:
- $df_1 = 2(numerator)$
- $df_2 = 6(denominator)$
- $IfF_{Cal} = 15.24 > F_{\text{critical}} = 5.14$ (or) $p < 0.05$, reject H₀.

Conclusion:

Since F ≈ 15.24 is likely greater than the critical value, we reject the null hypothesis and conclude that there is a significant difference in plant growth among the three fertilizers.

Source of Variation	**Degrees of Freedom (df)**	**Sum of Squares (SS)**	**Mean Square (MS)**	**F-Ratio**
Between Groups	44.19	2	22.095	15.24
Within Groups	8.68	6	1.447	
Total	52.87	8		

Interpretation:

- Since F = 15.24 is quite high and likely exceeds the critical F-value at α = 0.05, we reject the null hypothesis.
- This means at least one fertilizer has a significantly different effect on plant growth.
- Using the F-distribution table (or calculator), at α = 0.05, with df1 = 2 and df2 = 6, the critical value of F ≈ 5.14.

7.4. ANOVA TWO-WAY CLASSIFICATION WITH ONE OBSERVATION PER CELL:

ANOVA Two Way Classification

$Class\ i$	$Values\ Y_{ij}$ 1 2 ... n	$Total\ T_i$	$Mean$
1	$Y_{11}, Y_{12}, \dots, Y_{1n_1}$	T_1	$\bar{Y}_{1}.$
2	$Y_{21}, Y_{22}, \dots, Y_{2n_2}$	T_2	$\bar{Y}_{2}.$
…	…	…	…
i	$Y_{i1}, Y_{i2}, \dots, Y_{in_i}$	T_i	$\bar{Y}_{\imath}.$
…	…	…	…
k	$Y_{k1}, Y_{k2}, \dots, Y_{kn_k}$	T_k	$\bar{Y}_{k}.$
$Mean$	$Y_{.1}, Y_{.2}, \dots, Y_{.n_1}$	$Grand\ Total\ G$	$Grand\ Mean\ \bar{Y}_{..}$

Mathematical Model

$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

Where:

- Y_{ij} is j^{th} observation in the i^{th} sub classes

- $\mu = General\,Mean\,Effect$
- $\alpha_i = i^{th}\ sub\,class\,additive\,effect$
- $\beta_i = j^{th}$ sub class additive effect
- $\epsilon_{ij} \sim N(0, \sigma^2)$

Explanation of Terms:

- k: number of groups
- h= number of classes
- N: total number of observations
- G= Grand Total
- Set the Hypothesis
- Degrees of Freedom
- SS (Sum of Squares): a measure of variability
  - SST = total sum of squares
  - SSA = sum of squares Factor-A
  - SSB = sum of squares Factor-B
  - SSE =Sum of squares for Error
- MS (Mean Square): an average of the sum of squares (SS/df)
- F-ratio: used to determine statistical significance (if F is large enough, the null hypothesis is rejected)

a) Decomposition of Sum of Squares:

Step 1: Set Hypotheses

- $NullHypothesis(H_0): Allgroup\,means\,are\,equal$

  $H_{01}: \mu_1 = \mu_2 = \dots = \mu_k$

  $H_{02}: \mu_1 = \mu_2 = \dots = \mu_h$
- Alternative Hypothesis (H_1): At least one group mean differs

Step 2: Compute Group Totals and Means

- $i = 1, 2, \dots, k$
- $j = 1, 2, \dots, n_i$
- $\sum_{i=1}^{k} n_i \tau_i = 0(Constraintf\ or\ identifiability)$.
- k: number of groups

- N: total number of observations
- $Yij : Observation\ j\ in\ group\ i$
- $T_i = \sum_{j=1}^{n_i} Y_{ij} : Totalforgroup\ i$
- $\overline{Y_{\iota}}. = \frac{T_i}{n_i} : Meanforgroup\ i$
- $G = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} : Grandtotal$
- $\bar{Y} = \frac{G}{N}\ where\ N = \sum_{i=1}^{k} n_i : Grand\ mean$

Here unknown parameters are $\mu\ and\ \mathrm{t}_i$. We estimate the parameters by using the principle of least squares (method of least squares).

Minimize the error sum of squares partially differentiating w.r.t. the parameters.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad \text{-------------------}(1)$$

$$\epsilon_{ij} = Y_{ij} - \mu - \alpha_i - \beta_j$$

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} {\epsilon_{ij}}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu - \alpha_i - \beta_j\right)^2 - - - - - - (2)$$

Partial differentiation equ (2) w.r.t μ and equated to zero, we get

$$\frac{\partial E}{\partial \mu} = 0 \to 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu - \beta_j - \alpha_i\right)(-1) = 0$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$Y_{..} = kn_i\,\mu + n_i \sum_{i=1}^{k} \alpha_i + k \sum_{j=1}^{n_i} \beta_j \qquad \left(\therefore \sum_{i=1}^{k} \alpha_i = \sum_{j=1}^{n_i} \beta_j = 0\right)$$

$$\mu = \frac{\sum\sum Y_{ij}}{N} = \overline{Y_{..}} \qquad - - - - - - - - - - (3)$$

Partial differentiation equ (2) w.r.t α_i and equated to zero, we get

$$\frac{\partial E}{\partial \alpha_i} = 0 \to 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \mu - \alpha_i - \beta_j\right)(-1) = 0$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \alpha_i - \beta_j) = 0$$

$$\alpha_{i.} = \overline{Y_{i.}} - \overline{Y_{..}}$$

Partial differentiation equ (2) w.r.t β_j and equated to zero, we get

$$\alpha_{i.} = \overline{Y_{.j}} - \overline{Y_{..}}$$

$$\epsilon_{ij} = Y_{ij} - \mu - \alpha_i - \beta_j$$

$$\epsilon_{ij} = Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}}$$

Total Sum of Squares (SST):

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{..}})^2$$

Factor A Sum of Squares (SSA):

$$SSA = \sum_{i=1}^{k} n_i(\overline{Y_{i.}} - \overline{Y_{..}})^2$$

Factor B Sum of Squares (SSB):

$$SSB = \sum_{j=1}^{n_i} k(\overline{Y_{.j}} - \overline{Y_{..}})^2$$

Error Sum of Squares (SSE):

$$SSE = \sum(Y_{ij} - \overline{Y_{i.}} - \overline{Y_{.j}} + \overline{Y_{..}})^2$$

$$SSE = SST - SSA - SSB$$

Step 4: Compute Degrees of Freedom

$$df_{Factor\ A} = k - 1$$

$$df_{Factor\ B} = n - 1$$

$$df_{Error} = N - k - n - 1$$

$$df_{total} = N - 1$$

Step 5: Compute Mean Squares

$$MSA = \frac{SSA}{k-1}$$

$$MSB = \frac{SSB}{n-1}$$

$$MSE = \frac{SSW}{N-k-n-1}$$

Step 6: Compute F-Ratio

$$F = \frac{MSA}{MSE} \quad and \quad F = \frac{MSB}{MSE}$$

Step 7: Compare with F-Critical

$$Find\ F_{critical}\ from\ F - distribution\ table for given\ \alpha(usually\ 0.05), with$$

$$df_1 = k-1,\ df_2 = N-k$$

Decision Rule:

$$If\ F > F_{critical}: Reject H_0, \qquad If\ F \leq F_{critical}: Fail\ to\ reject H_0$$

Step 8: Conclusion

If H_0 is rejected, at least one group mean is significantly different.

$Source\ of\ Variation$	$Sumof\ Squares(SS)$	$Degreesof\ Freedom\ (df)$	$MeanSquare\ (MS) = SS/df$	$F - ratio - MS(Between)/MS(Within)$
Factor A	SSA	$k\ -\ 1$	$MSB_B = SSB_B/(k-1)$	$F\ = MSA/MSE$
Factor B	SSB	$h\ -\ 1$	$MSW_W = SSW_W/(h-1)$	$F\ = MSA/MSE$
Error	SSE	N-k-h+1		
$Total$	SST_T	$N\ -\ 1$		

a) Example for ANOVA Two way:

Sample (Factor A)	B1	B2	B3	B4	Row Total
S1	34	23	35	36	128
S2	33	36	32	35	136
S3	28	31	29	30	118
Column Total	95	90	96	101	G=382

Step 1: Set Hypotheses

$$NullHypothesis(H_0): Allgroup\ means\ are\ equal$$

$$H_{01}: \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_{02}: \mu_1 = \mu_2 = \ldots = \mu_h$$

Alternative Hypothesis (H_1): At least one group mean differs

Step 2: Calculate the Overall Mean (Grand Mean)

Grand total G:

G = 128 + 136 + 118 = 382

Number of rows r = 3, columns c = 4, total N=r×c=12

$$CF = \frac{G^2}{N} = \frac{(382)^2}{12} = 12{,}160.3333$$

Step 3: Compute the Sum of Squares (SST)

$$SST = \sum Y^2 - CF$$

$$\sum Y^2 == 34^2 + 23^2 + 35^2 + 36^2 + 33^2 + 36^2 + 32^2 + 35^2 + 28^2 + 31^2 + 29^2 + 30^2$$

$$= 1156 + 529 + 1225 + 1296 + 1089 + 1296 + 1024 + 1225 + 784 + 961 + 841 + 900 = 13{,}326$$

$$SST = 13{,}326 - 12{,}160.3333 = 1{,}165.6667$$

Sum of Squares for Factor A (Rows)

$$SSA = \frac{\sum(Row\ Total)^2}{C} - CF$$

$$= \frac{128^2 + 136^2 + 118^2}{4} - 12{,}160.3333$$

$$= \frac{16{,}384 + 18{,}496 + 13{,}924}{4} - 12{,}160.3333$$

$$= \frac{48{,}804}{4} - 12{,}160.3333$$

$$= 12{,}201 - 12{,}160.3333 = 40.6667$$

Sum of Squares for Factor B (Columns)

$$SSB = \frac{\sum(Column\,Total)^2}{r} - CF$$

$$= \frac{95^2 + 90^2 + 96^2 + 101^2}{3} - 12{,}160.3333$$

$$= \frac{9025 + 8100 + 9216 + 10201}{3} - 12{,}160.3333$$

$$= \frac{36{,}542}{3} - 12{,}160.3333$$

$$= 12{,}180.6667 - 12{,}160.3333 = 20.3334$$

Error (Residual) Sum of Squares:

Since we have no replication, the “error” is actually the interaction term

(Unexplained Variation):

$$SSE = SST - SSA - SSB$$

$$SSE = 1{,}165.6667 - 40.6667 - 20.3334 = 1{,}104.6666$$

Step 4: Degrees of Freedom

$$df_A = r - 1 = 2$$

$$df_B = c - 1 = 3$$

$$df_E = (r-1)(c-1) = 2 \times 3 = 6$$

$$df_T = N - 1 = 11$$

Step 5: Mean Squares & F-values

$$MS_A = \frac{SSA}{df_A} = \frac{40.6667}{2} = 20.33335$$

$$MS_B = \frac{SSB}{df_B} = \frac{20.3334}{3} = 6.7778$$

$$MS_E = \frac{SSE}{df_E} = \frac{1{,}104.6666}{6} = 184.1111$$

Step 6: Compute F-Ratio

$$F_A = \frac{MS_A}{MS_E} = \frac{20.33335}{184.1111} \approx 0.1104$$

$$F_B = \frac{MS_B}{MS_E} = \frac{6.7778}{184.1111} \approx 0.0368$$

Step 7: Compare with F-Critical

ANOVA Two-Way Table

Source	SS	df	MS	F
Factor A (Rows)	40.6667	2	20.33335	0.1104
Factor B (Cols)	20.3334	3	6.7778	0.0368
Error	1104.6666	6	184.1111	
Total	1165.6667	11		

Step 7: Conclusion

Interpretation:

Both $\mathrm{F_A}$ and $\mathrm{F_B}$ are far lessthan 1, so there is no statistically significant effect of either Sample (FactorA) or Column factor (FactorB). The large error term shows that most variation is unexplained.

7.5. BALANCED DESIGN AND UNBALANCED DESIGN:

Balanced Design:

A balanced design is one where:

- Each treatment (or factor level) has the same number of observations (replications).
- The data is evenly distributed across all groups or cells in the design.

Unbalanced Design an Unbalanced Design occurs when:

- The number of observations (replications) is not equal across treatment groups.
- Some cells (factor combinations) may even be missing entirely.

Feature	**Balanced Design**	**Unbalanced Design**
Replications per cell	Equal	Unequal / Missing
Analysis	Simple (standard ANOVA)	Complex (need Type I/II/III SS)
Power	Higher	Lower (if very uneven)
Interpretation	Easy	Sometimes tricky

7.6. CONCLUSION:

- One-Way ANOVA is used when a single factor with two or more levels is studied to check if there is any significant difference in the means of different groups. Example: comparing crop yields under different fertilizers.
- Two-Way ANOVA is applied when two factors are considered simultaneously. It evaluates:
  1) Main effect of factor A.
  2) Main effect of factor B.
  3) Interaction effect of A × B. Example: studying the effect of fertilizer type (Factor A) and irrigation level (Factor B) on crop yield.
- Assumptions of ANOVA:
  1) Observations are independent.
  2) Populations are normally distributed.
  3) Variances are equal across groups (homoscedasticity).
- ANOVA partitions total variation into between-group and within-group (error) variation. The F-test determines significance.
- Balanced designs (equal sample sizes) make ANOVA simpler, while unbalanced designs require advanced techniques.

7.7. SELF-ASSESSMENT QUESTIONS:

1) What is the purpose of ANOVA?
2) Differentiate between one-way and two-way ANOVA with examples.
3) Define main effect and interaction effect in two-way ANOVA.
4) What are the assumptions of ANOVA?
5) Write the mathematical model for one-way ANOVA.

6) Write the mathematical model for two-way ANOVA with interaction.

7) How is the F-ratio calculated in ANOVA?

8) Explain with an example where two-way ANOVA is more appropriate than one-way ANOVA.

9) What are degrees of freedom in one-way and two-way ANOVA?

10) State the difference between between-group and within-group variation.

7.8. SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.

2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

4) Hinkelmann, K., & Kempthorne, O. (2008). *Design and Analysis of Experiments*. Wiley.

5) Ott, R.L., & Longnecker, M. (2016). *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning.

6) Snedecor, G.W., & Cochran, W.G. (1989). *Statistical Methods*. Iowa State University Press.

Prof. V.V. Haragopal

LESSON-8

MULTIPLE COMPARISON TESTS

8.0. OBJECTIVES:

After completing this unit, you should be able to:

- Explain the need for multiple comparison tests after obtaining a significant ANOVA result.
- Distinguish clearly between Fisher's LSD and Duncan's Multiple Range Test based on their procedures and error-control strategies.
- Compare the advantages, limitations, and applications of the three methods.
- Use each test to determine which specific group means differ significantly in a dataset.
- Interpret the outcomes of multiple comparison tests and apply them effectively in real research situations.

STRUCTURE:

8.1. INTRODUCTION

In many scientific investigations-such as agricultural trials, medical experiments, psychological studies, or industrial quality testing-researchers often need to compare the performance of more than two groups or treatments. For example, a scientist may evaluate different fertilizers, a doctor may compare multiple drug dosages, or an educator may test various teaching methods. In all such situations, the central question is whether the **group means differ significantly**.

To address this, the Analysis of Variance (ANOVA) is commonly used as an initial statistical test. ANOVA examines whether there is **overall** variability among group means that cannot be explained by chance alone. If ANOVA yields a significant F-value, it tells us that **at least one** group mean is different from the others. However, ANOVA does **not** indicate:

- Which specific pairs of means differ,
- How large those differences are, or
- Whether the differences are practically meaningful.

This limitation requires additional statistical procedures known as **multiple comparison tests** or **post-hoc tests**.

Multiple comparison tests are specially designed to examine the **pairwise differences** among all group means in a controlled manner. Without these tests, comparing groups using several individual t-tests would greatly increase the likelihood of committing a **Type I error**, meaning we might incorrectly conclude that two groups differ when they do not. As the number of groups increases, the number of pairwise comparisons grows rapidly, and so does the risk of false positives.

To address this, multiple comparison procedures apply statistical corrections, adjustments, or decision rules that help maintain the overall accuracy of conclusions. These methods differ in their:

- Stringency or strictness (how much they control Type I error)
- Power (their ability to detect true differences)
- Assumptions (equal sample sizes, equal variances, normality) and
- Computational approaches.

Some methods, such as Fisher's Least Significant Difference (LSD), are more liberal and sensitive, making them good for detecting subtle differences but less strict in error control. Others, like Tukey's Honest Significant Difference (HSD), provide stronger protection against Type I error, especially with many groups. Duncan's Multiple Range Test lies between these methods, offering a balance of power and flexibility with a stepwise procedure.

Overall, multiple comparison tests play a crucial role in the interpretation of ANOVA results. They allow researchers to pinpoint exactly which means differ, understand the pattern of differences among treatments, and draw reliable conclusions from experimental data. By choosing an appropriate procedure based on the research design and objectives, investigators can ensure that their conclusions are both statistically sound and practically meaningful.

8.2. NEED FOR MULTIPLE COMPARISON TESTS:

A significant ANOVA result tells us that group differences exist, but it does not provide detailed information about where those differences lie. Conducting several independent t-tests is not recommended because it increases the probability of **Type I error** incorrectly concluding that differences exist when they do not. Multiple comparison tests offer a systematic and statistically valid solution by:

- Adjusting the significance level when multiple pairwise comparisons are made.
- Providing a consistent framework to determine which specific means differ.
- Protecting the study from inflated false-positive rates.
- Allowing researchers to make clear, interpretable decisions about treatment effectiveness or group behaviour.

Because different tests vary in strictness and power, selecting the right method ensures valid and reliable conclusions.

8.3. FISHER'S LEAST SIGNIFICANT DIFFERENCE (LSD) METHOD:

Fisher's LSD is one of the earliest and simplest post-ANOVA multiple comparison techniques. It operates by performing pairwise t-tests but only after a significant ANOVA result has been obtained. The method calculates a minimum difference-called the **Least Significant Difference**-that two means must exceed to be considered statistically different.

Key Features:

- It does not strongly adjust for multiple comparisons, making it **more liberal** (more likely to find differences).
- The method is powerful when the number of comparisons is small and when the overall ANOVA is highly significant.
- It uses the pooled variance from ANOVA to calculate the standard error for pairwise comparisons, which enhances consistency across tests.

1) Precondition:

- Perform one-way ANOVA first. Apply LSD **only if** the ANOVA F-test is significant at the chosen α (e.g. 0.05). LSD uses the pooled error variance (MSE) from that ANOVA.

2) Notation

- $\overline{X}_\iota$ and $\overline{X}_j$$: sample\ means\ of\ groups\ i\ and\ j$
- $n_i\ and\ n_j: sample\ sizes\ of\ groups\ i\ and\ j$
- MSE: $mean\ square\ error\ (pooled\ within - groups\ variance) from\ ANOVA$
- $df_e: error\ degrees\ of\ freedom\ from\ ANOVA$
- $t_{crit} = t_{\alpha/2,\, df_e}: two -$ $tailed\ critical\ t\ for\ significance\ level\ \alpha\ (commonly$ 0.05)
- LSD$_{ij}$$: least\ significant\ difference\ for\ comparison\ i vs\ j$
- $SE_{ij}: standard\ error\ of \overline{X}_\iota - \overline{X}_J$

3) Standard error formulas

Equal sample sizes (all groups have n)

$$SE_{ij} = \sqrt{\frac{2MSE}{n}}$$

Unequal sample sizes

$$SE_{ij} = \sqrt{MSE\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}$$

Note: LSD assumes homogeneity of variances (pooled MSE valid). If variances are unequal, LSD is not appropriate without modification.

4) LSD (Critical Difference):

$$LSD_{ij} = t_{\left(\frac{\alpha}{2}, df_e\right)} * SE_{ij}$$

- This is the minimum absolute difference between $\overline{X_i}\ and\ \overline{X_j}$ required for significance at the two-sided level α.

5) Decision Rule:

- $Compare\ \left|\overline{X_i} - \overline{X_j}\right|\ with\ \mathrm{LSD}_{ij}:$
- $If\left|\overline{X_i} - \overline{X_j}\right| > \mathrm{LSD}_{ij} \rightarrow \boldsymbol{significant}\ \left(reject\ H_0\!:\mu_i = \mu_j\right).$
- $If\ \left|\overline{X_i} - \overline{X_j}\right| \leq \mathrm{LSD}_{ij} \rightarrow \boldsymbol{notsignificant}.$

6) Equivalent Confidence Interval Form:

- A two-sided 100(1−α) % confidence interval for the difference $\mu_i - \mu_j$ is:
- $\left(\overline{X_i} - \overline{X_j}\right) \pm t_{\alpha/2,\, df_e}\, SE_{ij}.$
- If this interval **does not contain 0**, the difference is significant at level alpha α.

$\boldsymbol{Study}: Compare\ three\ fertilizers\ A, B, C.$

A	B	C
15.47	19.47	22.47
16.74	20.74	23.74
18.00	22.00	25.00
19.26	23.26	26.26
20.53	24.53	27.53

$Observed\ means\ and\ sample\ sizes:$

$$\overline{X_A} = 18, \quad \overline{X_B} = 22, \quad \overline{X_C} = 25, \qquad n_A = n_B = n_C = n = 5.$$

$From\ ANOVA\ (given)$:

MSE $= 4.0,\qquad df_{\text{error}} = 12.$

$ANOVA\ F\ was\ significant\ \rightarrow\ proceed\ with\ LSD\ at\ \alpha = 0.05$

$\boldsymbol{StepA - Formulaforstandarderror(equaln)}$

$$SE_{ij} = \sqrt{\frac{2 * MSE}{n}}$$

$\boldsymbol{Substitute}$:

$$SE = \sqrt{\frac{2 \times 4.05}{5}} = \sqrt{\frac{8}{5}} = \sqrt{1.6} = 1.2649110\ (\approx 1.265).$$

$\boldsymbol{Step\ B\ -\ Critical\ t}$

$Two - tailed\ ttt - critical\ with\ \alpha = 0.05\ and\ df\ = 12$:

$t_{crit} = t_{0.025,12} \approx 2.1788$ (often rounded to 2.179).

$\boldsymbol{StepC - LeastSignificantDifference(LSD)}$

LSD $= t_{crit} \times SE = 2.1788 \times 1.264911064 \approx 2.7560.$

$So\ any\ absolute\ mean\ difference\ > 2.7560\ is\ significant\ at\ \alpha = 0.05$

$\boldsymbol{Step\ D\ -\ Pairwise\ differences, compare\ with\ LSD}$

1. $B - A = 22 - 18 = 4.0\ > 2.7560 \Rightarrow\ \boldsymbol{significant}.$
2. $C - A = 25 - 18 = 7.0 > 2.7560 \Rightarrow\ \boldsymbol{significant}.$
3. $C - B = 25 - 22 = 3.0. > 2.7560 \Rightarrow\ \boldsymbol{significant}.$

$All\ three\ pairwise\ differences\ exceed\ the\ LSD$
$\rightarrow\ all\ means\ are\ significantly\ different.$

$\boldsymbol{StepE - Confidenceintervalsfordifferences(equivalentcheck)}$

$General\ CI$:

$$(\bar{X}_\iota - \bar{X}_J) \pm t_{crit} \cdot SE$$

- $For\ C - B$: $3.0 \pm 2.7560 = [0.2440,\ 5.7560] - interval\ does\ \textbf{not}\ contain0 \rightarrow significant.$
- $For\ B - A$: $4.0 \pm 2.7560 = [1.2440,\ 6.7560]$
- $For\ C - A$: $7.0 \pm 2.7560 = [4.2440,\ 9.7560]$

$\boldsymbol{Interpretation(Example1)}$

$At\ \alpha = 0.05\ C > B > A.$

$Fertilizer\ C\ gives\ the\ highest\ mean\ height, significantly\ higher\ than\ B\ and\ A;$

$B\ is\ significantly\ higher\ than\ A.$

Example:

Suppose we test the effect of 3 fertilizers (A, B, C) on plant growth.

- Fertilizer A: 20, 22, 23
- Fertilizer B: 25, 27, 26
- Fertilizer C: 22, 20, 21

From ANOVA, we get:

Means:

- $Fertilizer\ A = \frac{20+22+23}{3} = (21.7)$
- $Fertilizer\ B = \frac{25+27+26}{3} = (26.0)$
- $Fertilizer\ C = \frac{22+20+21}{3} = (21.0)$

Comparisons:

- $MSE\ =\ 1.0$
- $df_{error}\ = 6$
- $t_{0.05,6}\ = 2.447$
- $n\ =\ 3$

Fisher’s Least Significant Difference Formula:

$$LSD = t_{\left(1-\frac{\alpha}{2}, df_e\right)}\sqrt{\frac{2MSE}{n}}$$

$$LSD = 2.447 \times \sqrt{\frac{2 \times 1.0}{3}} = 2.0$$

Differences:

- A vs B = 4.33 → greater than 2.0 → significant
- A vs C = 0.67 → less than 2.0 → not significant
- B vs C = 5.0 → greater than 2.0 → significant

Conclusion:

Fertilizer B produces significantly more growth than A and C, but A and C do not differ.

8.4. DUNCAN'S MULTIPLE RANGE TEST (DMRT):

Duncan's Multiple Range Test is a post-hoc multiple comparison procedure used after ANOVA to determine which specific group means differ.

Developed by D.B. Duncan, it is considered less conservative than Tukey's HSD, meaning it is more likely to detect differences between groups.

It uses Studentized Range Statistics (q-values) but applies a *stepwise increasing significance level*, which gives DMRT more power (higher chance of finding differences).

Key Features of DMRT:

- Stepwise procedure - comparisons begin with the largest range (largest difference between means).
- Uses q-statistics from the Studentized Range distribution.
- More liberal than Tukey's HSD, but less liberal than unadjusted LSD.
- Controls Type I error at each step but not the experiment-wise error**.**

Steps in Duncan's Multiple Range Test

1) Perform ANOVA DMRT is used only if the ANOVA F-test is significant.
2) Arrange means in ascending or descending order
3) Compute the Standard Error (SE)

$$\boldsymbol{SE} = \sqrt{\frac{MSE}{n}}$$

where

- MSE = Mean Square Error from ANOVA
- n = number of observations per group (for equal sample size)

1) Find the least significant ranges (LSR)
For a range of r means:

$$LSR(r) = q_{r,\, df\, error} \times SE$$

2) **Compare differences between ordered means against LSR values**
If the difference is greater than LSR → means are significantly different.

Example:

Fertilizer	**Plant 1**	**Plant 2**	**Plant 3**	**Mean**
A	20	22	23	21.7
B	25	27	26	26.0
C	22	20	21	21.0

From **ANOVA**, suppose:

$\boldsymbol{Step\ 1\ -\ Order\ means}$

$$C = \frac{22 + 20 + 21}{3} = (21.0)$$

$$A = \frac{20 + 22 + 23}{3} = (21.7)$$

$$B = \frac{25 + 27 + 26}{3} = (26.0)$$

Comparisons

- $MSE\ =\ 1.56$
- $Error\ df\ =\ 6$

$\boldsymbol{Step\ 2\ -\ Calculation}$

$$SE = \sqrt{\frac{1.56}{3}} = 0.72$$

$\boldsymbol{Step\ 3\ -\ Critical\ q - values}$

$Take\ from\ Studentized\ Range\ Table(dependsondf, \alpha = 0.05).$

$For\ r = 2, q_2 = 2.95$

$For\ r = 3, q_3 = 3.31$

$\boldsymbol{Step\ 4\ -\ Compute\ LSRs}$

- $LSR_2 = 2.95 \times 0.72 = 2.12$
- $LSR_3 = 3.31 \times 0.72 = 2.38$

Step 5 – Compare Differences

- B vs C = 26.0 – 21.0 = 5.0 > 2.38 → Significant
- B vs A = 26.0 – 21.7 = 4.3 > 2.12 → Significant
- A vs C = 21.7 – 21.0 = 0.7 < 2.12 → Not significant

Conclusion: Fertilizer B is significantly better than A and C, but A and C are similar.

8.5. DIFFERENCE BETWEEN FISHERS AND DUNCANS:

Feature	Fisher's LSD	Duncan's Multiple Range Test (DMRT)
Type of method	Pairwise comparison using pooled t-tests	Stepwise multiple range test
Protection against Type I error	**Weak control** - higher chance of false positives	**Moderate control**, stronger than LSD but weaker than Tukey
Requires significant ANOVA first?	Usually **yes** (Fisher's rule)**	**Yes**, but still conducts stepwise comparisons
Basis of critical value	**Constant critical difference (LSD)** using t-value for all comparisons	**Variable critical ranges (R values)** depending on the number of ordered groups compared
Comparison approach	Compares all pairs equally	Compares **ordered means** in a step-down procedure
Stringency	More liberal (detects more differences)	More conservative than LSD, but liberal compared to Tukey
Risk of Type I error	High	Medium
Power (ability to detect real differences)	High (but risks false alarms)	Medium-high
Best used when	Few groups + low risk of false positives is acceptable	Agricultural / biological experiments with ordered treatments
Output style	Pairwise tests with single LSD value	Means are grouped into **homogeneous subsets** (e.g., a, b, c letters)
Interpretation	"Means differ if	Xi – Xj

8.6. APPLICATIONS:

- **Agriculture:** Comparing crop yields under different fertilizers.
- **Medicine:** Comparing effects of different drug dosages.
- **Education:** Comparing student performance under different teaching methods.
- **Psychology / Behavioral Science:** Comparing stress levels under different relaxation techniques (e.g., meditation, music therapy, exercise).
- **Manufacturing / Industry:** Comparing the strength of materials produced by different production processes.

8.7. SUMMARY OF MULTIPLE COMPARISON PROCEDURES:

Fisher's Least Significant Difference (LSD) test is one of the earliest and simplest post-hoc methods used after ANOVA to identify which group means differ significantly. It compares pairs of means using the pooled error variance from ANOVA and relies on the *t*-distribution. Because it does not strongly control the familywise Type I error rate, it is considered a **liberal** method-meaning it often detects significant differences, but at the cost of a higher risk of false positives. Fisher's LSD is most suitable when the number of treatments is small and when researchers want a highly sensitive method to detect differences.

Duncan's Multiple Range Test (DMRT) is a stepwise procedure that uses the studentized range statistic (*q*) and compares ordered means to determine significant differences. It provides better error control than LSD while still remaining more powerful than conservative tests like Tukey's HSD. DMRT groups means into homogeneous subsets (A, B, AB, etc.) based on their statistical similarity, making interpretation easier in agricultural and biological experiments. Although DMRT is less liberal than LSD, it still allows more flexibility than stricter methods, balancing sensitivity and protection against false positives.

8.8. SELF-ASSESSMENT QUESTIONS:

1) What is Fisher's LSD test used for after ANOVA?
2) Why is Duncan's test considered less strict than other multiple comparison tests?
3) Differences Between Fishers LSD and Duncan's Test?
4) How does Duncan's test group treatments compare to Fisher's LSD?
5) Problems on LSD and Duncans Test?

8.9. SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

4) Hinkelmann, K., & Kempthorne, O. (2008). *Design and Analysis of Experiments*. Wiley.

5) Ott, R.L., & Longnecker, M. (2016). *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning.

6) Snedecor, G.W., & Cochran, W.G. (1989). *Statistical Methods*. Iowa State University Press.

Dr. M. Amulya

LESSON-9

FIXED, RANDOM AND MIXED EFFECT MODELS

9.0. OBJECTIVES:

After Studying this unit, you should able to:

- Understand the difference between fixed, random, and mixed effect models.
- Identify situations where each model is applicable.
- Learn the assumptions behind each model.
- Apply these models in practical research problems.
- Compare their advantages and limitations

STRUCTURE:

9.1. INTRODUCTION

In statistical modeling and analysis of variance, factors influencing a response can be treated as **fixed, random, or mixed effects** depending on how their levels are chosen. In a **fixed effect model**, the levels of the factor are specifically selected by the researcher, and inference is restricted to those levels only. In contrast, a **random effect model** assumes that the factor levels are randomly drawn from a larger population, allowing generalization beyond the sample. A **mixed effect model** combines both, where some factors are fixed and others are random, making it suitable for more complex designs. These models are fundamental in agriculture, medicine, engineering, and social sciences for designing experiments and interpreting results accurately.

9.2. FIXED EFFECT MODEL:

A fixed effect model is used when the levels of a factor are specifically chosen by the **researcher** and are the only ones of interest. The purpose is to compare these selected treatments without generalizing beyond them. The treatment effects are considered constant (non-random), and inference is limited to the chosen levels.

Model Equation:

$$\mathbf{Y}_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\,(0, \sigma^2)$$

Where:

- $\mathbf{Y}_{ij}$= Outcome for group i at time t
- μ = overall mean
- α_i = fixed treatment effect of i^th level

- ε_{ij} = random error associated with i^th level treatment

Example:

A researcher wants to test whether average exam scores differ across three teaching methods (A, B, C). The data are given below:

Method	S1	S2	S3	S4
A	78	74	82	80
B	85	88	90	87
C	72	70	68	69

Solution Steps:

1) **Model:** $Yij \ = \ \mu \ + \ \tau_i \ + \ \varepsilon_{ij}$

   where τ_i are fixed effects of teaching method.

2) **Hypotheses:**

   $H_0: \tau_a \ = \tau_b \ = \tau_c \ = \ 0 \ (no\ difference)$

   H_a: At least one $\tau_i \neq 0$

3) **Means:**

   Method A mean = 78.50

   Method B mean = 87.50

   Method C mean = 69.75

   Grand mean G = 78.58

4) **Calculate Sum of Squares**

   - Total Sum of Squares (SST):

$$SST_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2 = 686.92$$

- Between Groups Sum of Squares (SSB):

$$SSB_B = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2 = 630.17$$

- Within Groups Sum of Squares (SSW):

$$SSW_W = SST_T - SSB_B = 56.75$$

5) **Degrees of Freedom:**

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

$$df_{within} = N - k = 12 - 3 = 9$$

$$df_{total} = N - 1 = 12 - 1 = 11$$

6) **Mean Squares:**

$$MSB_B = \frac{SSB}{k-1} = 315.08$$

$$MSW_W = \frac{SSW}{N-k} = 6.31$$

7) **F-ratio:**

$$F = \frac{MSB_B}{MSW_W} = \frac{315.08}{6.31} \approx 49.97$$

8) **Decision:**

Critical F (2,9) at α=0.05 ≈ 4.26. Since 49.97 >> 4.26, reject H_0.

Conclusion:

Teaching methods have a significant effect on exam scores. Method B performs best, Method A is average, and Method C performs worst.

9.3. RANDOM EFFECT MODEL:

A random effect model is applied when the factor levels are randomly sampled from a larger population. Here, the focus is not on comparing specific treatments but on estimating the variability among treatments. The treatment effects are assumed to be random variables with mean zero and constant variance.

Model Equation:

$Y_{\mathrm{ij}} = \mu + \alpha_i + \epsilon_{ij}, \quad \alpha_{\mathrm{i}} \sim N(0, \sigma^2\alpha), \quad \varepsilon_{\mathrm{ij}} \sim N(0, \sigma^2)$

Where:

- $\mu = overall\,mean$
- $\alpha_i = random\,effect\,of\,the\,ith\,level\,treatment,$
- ϵ_{ij} = random error

Example:

A manufacturer wants to estimate variability in product weight due to machines. Four machines are randomly selected from the factory floor and each machine produces 3 items. The weights (in grams) recorded are:

Machine	Item 1	Item 2	Item 3
M1	50.2	49.8	50.5
M2	51.0	50.6	50.9
M3	49.0	48.7	49.3
M4	50.7	50.4	50.8

Model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

$\alpha_i \sim N(0, \sigma_\alpha^2)(randommachineeffect)$

$\varepsilon_i j \sim N(0, \sigma^2)(within - machineerror)$

$i = 1, \ldots, k\ machines; \ j = 1, \ldots, n\ items\ per\ machine.$

$1)\ Compute\ machine\ (cell)\ means\ and\ grand\ mean:$

$M1\ mean = (50.2 + 49.8 + 50.5)/3 = 50.1667$

$M2\ mean = (51.0 + 50.6 + 50.9)/3 = 50.8333$

$M3\ mean = (49.0 + 48.7 + 49.3)/3 = 49.0000$

$M4\ mean = (50.7 + 50.4 + 50.8)/3 = 50.6333$

$Grand\ mean = (sumofall\ 12\ observations)/12 = 50.1583$

$2)$ ***Sums of Squares***

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \hat{y}Y_{..})^2 = 6.8246$$

$$SSB = n\sum_{i=1}^{k}(\hat{y}Y_{i.} - \hat{y}Y_{..})^2 = 5.9300$$

$$SSE \ = \ SST \ - \ SSB = 0.8946$$

$3)\ Degrees\ of\ freedom:$

$$df_B = k - 1 = 3$$

$$df_E = k(n-1) = 8$$

$$df_T = N - 1 = 11$$

$4)$ ***Mean Squares***

$$MSB = SSB/(k-1) = 5.9300/3 = 1.9767$$

$$MSE = SSE/[k(n-1)] = 0.8946/8 = 0.1118$$

$5)\ Variance - component\ estimates:$

$$\hat{\sigma}^2 = MSE = 0.1118$$

$$\hat{\sigma}_\alpha^2 = (MSB - MSE)/n = (1.9767 - 0.1118)/3 = 0.6216$$

$6)\ Intraclass\ Correlation\ (ICC):$

$$ICC = \hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}^2)$$

$$ICC = 0.6216/(0.6216 + 0.1118) = 0.847\ (\approx 84.7\%)$$

$Interpretation: about\ 84.7\%\ of\ the\ total\ variance\ is\ due\ to\ differences\ between\ machines.$

$7)$ ***F − test for random effect***

$$F = \frac{MSB}{MSE} = \frac{1.9767}{0.1118} \ = \ 17.69\ with\ df1 = k-1 = 3\ and\ df2 = k(n-1) = 8$$

$Critical\ F(3{,}8)\ at\ \alpha = 0.05 \approx 4.07. Since\ 17.69$
$> 4.07, machine\ effects\ are\ significant.$

ANOVA Summary Table

Source	*SS*	*df*	*MS*	*F*
Machines	5.9300	3	1.9767	17.69
Error	0.8946	8	0.1118	
Total	6.8246	11		

Conclusion:

The random effect of machine is significant (F = 17.69, $p < 0.05$). Estimated variance components: between-machine variance ≈ 0.6216, within-machine variance ≈ 0.1118. High ICC (≈0.85) indicates most variability comes from machine-to-machine differences.

9.4. MIXED EFFECT MODEL:

A mixed effect model includes **both fixed and random factors**. Some effects are chosen deliberately (fixed), while others represent random variation. These models are useful when experiments involve structured treatments combined with naturally occurring random factors, such as blocks or subjects.

Model Equation:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \qquad fixed, \quad \beta_j \sim N(0, \sigma^2\beta)$$

μ= overall mean

α_i= fixed effect of the ith treatment

β_j = random effect of the jth block

ε_{ijk} = random error

Problem:

A manufacturer wants to estimate variability in product weight due to machines. Four machines (random sample) are selected and each receives two fertilizers (fixed: F1, F2). Each machine–fertilizer combination is observed once (one measurement per cell). The weights (in grams) are:

Machine	**F1**	**F2**
M1	52	55
M2	48	50
M3	51	53
M4	49	52

$Machine\ means$:

$M1\ mean$: 53.50,

$M2\ mean$: 49.00

$M3mean$: 52.00,

$M4\ mean$: 50.50

$Grand\ mean\ \bar{Y}.. =\ 51.25$

$Sums\ of\ Squares$

$SST = \Sigma_i \Sigma_j (Y_i j - \bar{Y}..)^2 = 35.5000$

$SSB = a\Sigma_i (\bar{Y}_i. - \bar{Y}..)^2 = 22.5000$

$SSA = b\Sigma_j \left(\bar{Y}_{.j} - \bar{Y}..\right)^2 = 57.5000$

$SSAB = SST - SSA - SSB = 0.5000$

Mean Sum of Squares

$MSA = SSA/(a-1)$,

$MSB = SSB/(b-1)$,

$MSAB = SSAB/[(a-1)(b-1)]$

$F_t test\ for\ Fertilizer\ (fixed): F = MSA/MSAB, \big(df1 = a-1, df2 = (a-1)(b-1)\big)$

$F_t est\ for\ Machines\ (random): F = MSB/MSAB, \big(df1 = b-1, df2 = (a-1)(b-1)\big)$

$$EMS: E[MSA] = \sigma^2 + b\sigma_{\alpha\beta}^2 + br * (fixedeffectterm), E[MSB] = \sigma^2 + a\sigma_{\alpha\beta}^2 + ar\sigma_\beta^2, E[MSAB] = \sigma^2 + r\sigma_{\alpha\beta}^2$$

$Becauser\ = 1, \sigma^2 and \sigma_{\alpha\beta}^2 arenotseparatelyidentifiable; MSABestimatestheirsum.$

9.5. COMPARISON OF MODELS:

Aspect	Fixed Effect Model	Random Effect Model	Mixed Effect Model
Definition	Specific treatments chosen by researcher.	Treatments are random samples from a population.	Combination of fixed and random factors.
Inference Scope	Limited to chosen treatments only.	Generalizes to the entire population.	Both specific and general conclusions.
Treatment Effect	Constants (τi)	Random variables (αi)	Mix of constants and random variables.
Estimation Focus	Differences between means.	Variance components.	Both mean differences and variance components.
Examples	Fertilizers A, B, C.	Randomly chosen schools or machines.	Fertilizers (fixed) + fields (random).
Advantages	Simple, easy to interpret.	Allows broad generalization.	Handles complex, realistic designs.
Disadvantages	No generalization possible.	Complex estimation.	Computationally intensive.

9.6. APPLICATIONS:

1) **Agriculture** – Fertilizer trials (fixed), soil plots (random), mixed designs for crop yield.
2) **Medicine** – Drug dosage levels (fixed), patient-to-patient variation (random).
3) **Industry** – Comparing manufacturing methods (fixed), machine variation (random).
4) **Education** – Studying teaching methods (fixed), schools or classrooms (random).
5) **Psychology** – Comparing therapy types (fixed) while accounting for subject variability (random).
6) **Environmental Studies** – Pollution control methods (fixed) tested across random locations (random).

9.7. SUMMARY:

Fixed models are used when the levels of a factor (treatments, groups, or categories) are specifically chosen and represent the *entire set of interest*. In other words, the researcher wants to draw conclusions only about the treatments included in the study. Because of this, differences detected among treatment means apply exclusively to those particular levels.

Fixed models are common in agricultural experiments, lab studies, and clinical trials where treatments such as fertilizer types, drug doses, or teaching methods are deliberately selected. In fixed-effects ANOVA, the treatment means are compared directly, and statistical tests focus on identifying specific differences among them.

Random models, by contrast, consider treatment levels as a *random sample from a much larger population of possible levels*. The goal is not to study those levels individually but to generalize findings to the broader population. Instead of testing differences among specific means, random-effect models focus on estimating variance components-how much variability in the response is due to the random factor. This is useful in biological studies, multi-site experiments, and situations with subjects drawn randomly from populations.

Mixed models combine both fixed and random effects, allowing some factors to be specific (fixed) and others to represent random variability. These models are essential in designs like **Randomized Block Designs (RBD)**, **Latin Square Designs (LSD)**, repeated-measures studies, and multi-level data where blocks, subjects, or locations act as random factors. Mixed models improve precision by accounting for structured random variation, enabling both specific treatment comparisons and broader generalization.

9.8. SELF-ASSESSMENT QUESTIONS:

1) Differentiate between fixed, random, and mixed effect models with examples.
2) Write the assumptions of the random effect model.
3) Why are mixed models useful in agricultural and industrial research?
4) List advantages and disadvantages of fixed effect models.
5) Give one real-life situation for each type of model.
6) In which field would you apply mixed effect models and why?

9.9 SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
2) Hinkelmann, K., & Kempthorne, O. (2008). *Design and Analysis of Experiments*. Wiley.
3) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Dr. M. Amulya

LESSON-10

ADVANCED ANALYSIS OF VARIANCE

10.0. OBJECTIVES:

After reading this unit

- Become familiar to the analysis of variance technique
- Describe the various types of analysis of variance technique
- Describe the various types of assumptions involved in analysis of variance
- Define the various types of linear models used in analysis of various
- Describe the applications of analysis of variance

STRUCTURE:

10.1. INTRODUCTION:

Analysis of variance (ANOVA) was developed by the pioneering British Statistician Sir Ronald Aylmer Fisher (R.A. Fisher), who introduced the technique in the 1920s, notably in his book statistical methods for Research Workers (1925) and The Design of Experiments (1935) for analysing agricultural data, allowing researchers to Compare means of multiple groups by partitioning total variance

10.2. CONCEPT OF ANOVA:

ANOVA is a statistical method that analyzes variances to determine if the means from more than two populations are same. In other words, we have a quantitative response variable and a categorical explanatory variable with more than two levels. In ANOVA, the categorical explanatory is typically referred as the factor.

Analysis of Variance (ANOVA) is a parametric statistical technique used to compare datasets. This technique was developed by R.A. Fisher, and hence it is refereed in his name as Fisher's ANOVA. Its applications are similar to other statistical techniques such as t-test and z-test and this test is applied for comparing the means and relative variance between samples or population. ANOVA is considered as paramount test to compare more than two populations or samples.

This parametric statistical technique holds certain important assumptions including the following:

1) **Independence of Case:** In this assumption the dependent variable should be independent or randomly sample should be selected without any pattern.

2) **Normality:** In this, the assumption followed is that each group should be distributed normal. The normality of the group is confirmed by carrying out tests like Kolmogorov-Smirnov or the Shapiro-Wilk test.

3) **Homogeneity:** If the means variance between the groups is same, then it is called as Homogeneity. It can be tested using Levene's test.

If any data follow the above assumptions, then the analysis of variance (ANOVA) would be the appropriate technique for carrying out the comparison between the means of two, or more, populations.

10.3. TERMINOLOGIES RELATED TO ANOVA:

10.3.1. Sum of Square between Groups:

For the sum of the square between groups, the individual means of the group are calculated followed by the deviation from the individual mean for each group are taken. Finally sum of all groups is taken. It is also called as 'between groups variance' and denoted as SS(B).

10.3.2. Sum of Squares within Groups:

For sum of squares within a group, first, the grand mean for all groups are calculated and deviation from the individual group is taken. Finally, the sum of all groups will be made after squaring the deviation. It is also called as 'within group's variance' and denoted as SS(W).

10.3.3. F-Ratio:

It is calculated by dividing the sum of the squares between groups by the sum of the square within a group.

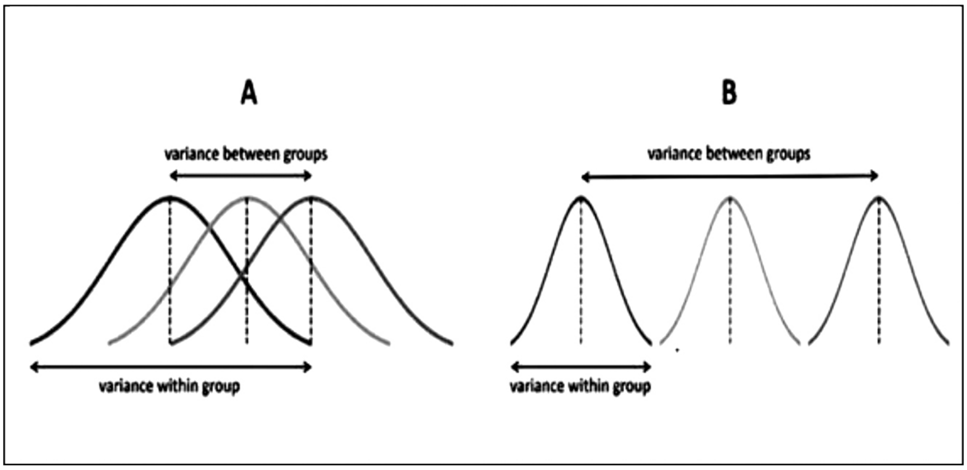

Fig. 10.1: Graphical Representation of Analysis of Variance

From the above graphical representation, Figure ‘A’ reveals that variance is large within group while it is small in the case of between the groups. Thus, the calculated ‘F’ value will be smaller. It indicates that there is no significant difference between groups. In contrast, Fig ‘B’ reveals that variance is small within groups but larger difference in variance is observed between groups. It thus interprets that there is significance difference between groups.

10.3.4. Degree of Freedom:

Degree of freedom (DF) for sum of square between group (SS(B)) is calculated by deducting value one (1) from the number of samples groups (k). Hence it is denoted as DF is k-1. In the case of sum of squares within group (SS(W)), the degree of freedom is calculated by deducting number of sample groups (k) from the total observation (N). Thus, the DF is denoted as N-k.

10.3.5. Significance:

It is important component of ANOVA where level of significance plays an important role in acceptance or rejection of hypothesis or null hypothesis respectively. Generally, it is defined as the probability of rejecting the null hypothesis when it is true at a predetermined level of significance say, 5%, 1%.

Generally, two ways of comparison for significant of ANOVA, that is based on F - value and P – value.

1) If calculated significance value (F) is compared with critical table value (i.e F- distribution table value);

    - If calculated F value is less than the Critical F value, we accept the null hypothesis, and then it is interpreted as there is no difference between the groups means.

    - If calculated F value is greater than the Critical F value, we reject the nullhypothesis, and then it is interpreted as there is difference between the groups means.

2) If calculated probability significance value (p) is compared with predetermined level of significance value (usually at 5%);

   - If ‘p’ value is smaller than the predetermined significance level value, we reject null hypothesis, and then it is interpreted as there is difference between the group means.
   - If ‘p’ value is greater than the predetermined significance level, we accept null hypothesis, and then it can be interpreted as there is no significant difference between groups.

Nowadays modern computers can automatically calculate the probability value for F- ratio.

10.4. TYPES OF ANOVA:

There are three types of ANOVA.

1) One -Way Analysis
2) Two -Way Analysis
3) K-Way Analysis

In these three types of analysis mainly we are using two types that are one way and two-way analysis of variances.

10.4.1 One-Way Analysis:

When we are comparing more than three groups based on one factor variable, then it is said to be one-way analysis of variance (ANOVA). One-Way ANOVA is a parametric test. This test is also known as One-Factor ANOVA / One-Way Analysis of Variance / Between Subjects ANOVA.

Statistical Analysis of the Model:

Let us suppose that N observations $X_{ij}(\, i = 1, 2, ...,k; j = 1, 2, ..., r)$ of a random arableX are grouped, on some basis, into k classes of sizes n_1, n_2, ..., n_k respectively, $\left(N = \sum_{i=1}^{k} n_i \right)$ as exhibited below:

	1	**2**	**….**	**j…**	**R**	**Total**	**Mean**
1.	X_{11}	X_{12}	. . . .	X_{1j} . . .	$X_{1r.}$	$X_{1.}$	$\overline{X}_{1}.$
2.	X_{21}	X_{22}	. . . .	X_{2j} . . .	$X_{2r.}$	$X_{2.}$	$\overline{X}_{2}.$
. . .	. . .	. . .	… … …	. . .	. . .	. . .	. . .
i	X_{i1}	X_{i2}	. . . .	X_{ij} . . .	X_{ir}	$X_{i}.$	$\overline{X}_{i}.$
. . .	. . .	. . .	… … …	. . .	. . .	. . .	. . .
k	X_{k1}	X_{k2}	. . . .	X_{kj} . . .	X_{kr}	$X_{k}.$	$\overline{X}_{k}.$

Mathematical Model:

Let xij be the Individual measurement of jth experimental units for ith treatment. The mathematical model for one-way classification is

$$x_{ij} = \mu + \alpha_i + e_{ij} \ \forall i = 1.2,...k, \ j = 1,2,...r \qquad [1]$$

where, μ = General mean

α_i = ith treatment class effect

e_{ij} = Random error

$$x_{ij} \sim N\left(\mu + \alpha_i, \sigma^2\right)$$

Here e_{ijk} is random errors which are *identically and independently distributed (iid)* following N (0, σ^2).

Assumptions in the Model:

1) All the observations are independent
2) Deferent effects are additive in nature.
3) $e_{ij} \sim \text{idd}\, N\left(0, \sigma^2\right)$

Null Hypothesis:

In a one-way ANOVA, there are two possible hypotheses. Let us consider the null hypothesis under consideration is

$$H_0 : \quad \mu_1 = \mu_2 = ... = \mu_k \ \mu$$

(or)

$$H_0 : \quad \alpha_1 = \alpha_2 = ... = \alpha_k \ 0$$

(or)

H0: There is no significance difference between the treatments.

To test the above Null hypothesis, first we estimate the parameters in mathematical model (1) by using the principle of least square by minimizing the error sum of squares. By solving the equation (1), we get the following results.

Grand Total: $G = \sum_{i=1}^{k} \sum_{j=1}^{r} x_{ij}$ Correction Factor: $CF = \frac{G^2}{rk}$ Since N = r * k

Total sum of squares: $TSS = \sum_{i=1}^{k} \sum_{j=1}^{r} x_{ij}^2 - CF$

Sum of squares due to treatment: $SSTr = \frac{\sum_{i=1}^{r} x_i^2}{r} - CF$

Sum of squares due to Error: SSE = TSS – SSTr

Degrees of Freedom:

Degrees of freedom carried by TSS is (rk - 1)

Degrees of freedom carried by SSTr is (k- 1)

Degree of freedom carried by SSE is k(r - 1)

ANOVA Table:

To the above null hypothesis by using this calculation, we construct the following ANOVA Table.

Source of Variation	Degrees of freedom	Sum of squares	Mean Sum of Squares	F-Ratio	
				F cal.val	F cri.val
Treatments	k-1	SST_r	$MSS_{T_r} = \frac{SSTr}{k-1}$	$F_{Tr} = \frac{MSS_{Tr}}{MSS_E}$	$F[k-1, k(r-1)$ @ α% los
Error	$k(r-1)$	SSE	$MSS_E = \frac{SSE}{k(r-1)}$	-	-
Total	rk-1	TSS		-	-

Statistical Decision:

We compare F calculated value with F critical values @ 5% los. We draw the conclusions accordingly.

10.4.2. Two-Way Analysis:

The two-way analysis of variance is an extension to the one-way analysis of variance. When factor variables are more than two, then it is said to be two-way analysis of variance (ANOVA). That is, when the data is classified into groups according to only two factors, like age group and gender we call it a two – way classified data and the corresponding ANOVA is called the Two-way ANOVA. At each combination of the levels of the factors, there may be more than one data value. This is called *replication*. Two-way tests can be with or without replication.

- **Two-Way ANOVA with Replication:** When there are replications, it is possible to estimate the interaction or the joint effect of the two factors on the response being studied.
- **Two-Way ANOVA without Replication:** When there are no replications, we can still perform two-way ANOVA. In this case, interactions cannot be estimated.

Statistical Analysis of the Model:

Let there be an 'N' experimental unit, in this experiment 'k' is number of treatments and 'r' is number of blocks. Here the total variation is divided into three parts.

1) Variation between the Treatments
2) Variation between Blocks
3) Variation due to Error

If there are 'r' such blocks, we say that the blocks are at 'r' levels. Similarly, if there are 'k' treatments, we say that the treatments are at 'k' levels. The responses from the 'r' levels of blocks and 'k' levels of treatments can be arranged in a two-way layout. The observed data set is arranged as follows:

Blocks **Treatments**	**1**	**2**	**. . . .**	**j . . . .**	**r**	**Block Total**	**Block Mean**
1.	X_{11}	X_{12}	. . . .	X_{1j} . . .	X_{1r}	$X_{1}.$	$\overline{X}_{1}.$
2.	X_{21}	X_{22}	. . . .	X_{2j} . . .	X_{2r}	$X_{2}.$	$\overline{X}_{2}.$
. . .	. . .	. . .	… … …	. . .	. . .	. . .	. . .
K	X_{i1}	X_{i2}	. . . .	X_{ij} . . .	X_{ir}	$X_{i}.$	$\overline{X}_{k}.$
Treatment Total	$X_{\cdot 1}$	$X_{\cdot 2}$	. . . .	$X_{\cdot j}\ldots$	$X_{\cdot r}$	$X_{k}.$ **Grand Total**	$\overline{X}_{k}.$ **Grand Mean**
Treatment Mean	$\overline{X}_{\cdot 1}$	$\overline{X}_{\cdot 2}$	. . . .	$\overline{X}_{\cdot j}\ldots$	$\overline{X}_{\cdot r}$		

Mathematical Model:

Let x_{ij} be the yield from i^{th} treatment and j^{th} block. The mathematical for two-way classification as follows

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \text{ for all } i = 1, 2, \ldots k,\ j = 1, 2, \ldots r$$

where, μ = General mean

α_i = i^{th} treatment effect

β_j = j^{th} block effect

e_{ij} = Random error

$e_{ij} \sim N(0, \sigma^2)$

Null Hypothesis:

There are three pairs of null or alternative hypotheses for the two-way ANOVA. Let us consider the null hypothesis under consideration is

1) H_0: All treatments are homogeneous

(or)

$$H_0: \alpha_1 = \alpha_2 = \ldots.. = \alpha_k = 0 \text{ (or) } \sum_{i=1}^{k} \alpha_i = 0$$

2) H0: All blocks are homogeneous

(or)

$$H_0: \beta_1 = \beta_2 = ... = \beta_r = 0 \text{ (or) } \sum_{i=1}^{r} \beta_j = 0$$

To test the above Null hypothesis, first we estimate the parameters in mathematical model (1) by using the principle of least square by minimizing the error sum of squares.

By solving the equation (1), we get the following results.

Grand Total: $G = \sum_{i=1}^{k} \sum_{j=1}^{r} x_{ij}$

Correction Factor: $CF = \frac{G^2}{rk}$ Since N = r * k

Total sum of squares: $TSS = \sum_{i=1}^{k} \sum_{j=1}^{r} x_{ij}^2 - CF$

Sum of squares due to treatment: $SSTr = \frac{\sum_{i=1}^{k} x_i^2}{r} - CF$

Sum of squares due to Block: $SSB = \frac{\sum_{j=1}^{r} x_j^2}{k} - CF$

Sum of squares due to Error: SSE = TSS – SSB – SSTr

Degrees of Freedom:

Degrees of freedom carried by TSS is (rk – 1)

Degrees of freedom carried by SSTr is (k – 1)

Degrees of freedom carried by SSB is (r – 1)

Degree of freedom carried by SSE is (k – 1) (r – 1)

ANOVA Table:

To the above null hypothesis by using this calculation, we construct the following ANOVA Table.

Source of Variation	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-Ratio	
				F cal.val	F cri.val
Treatments	k-1	SST$_r$	$\text{MSS}_{T_r} = \dfrac{\text{SSTr}}{k-1}$	$F_{Tr} = \dfrac{\text{MSS}_{Tr}}{\text{MSS}_E}$	$F[k-1, k(r-1)$ $@\ \alpha\%$ los
Blocks	$r-1$	SSB	$\text{MSS} = \dfrac{\text{SSB}}{r-1}$	$F_{Tr} = \dfrac{\text{MSB}}{\text{MSS}_E}$	$F[r-1,(k-1)(r-1)]\ @\ \alpha\%$ los
Error	$k(r-1)$	SSE	$\text{MSS}_E = \dfrac{\text{SSE}}{(k-1)(r-1)}$	-	-
Total	rk-1	TSS		-	-

Statistical Decision:

We compare 'F' calculated value with 'F' critical values @ 5% LOS. We draw the conclusions accordingly.

10.4.3. Comparison between One Way and Two-Way ANOVA:

Basis for Comparison	**One Way ANOVA**	**Two Way ANOVA**
Meaning	One-way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance	Two ways ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Compares	Three or more levels of one factor	Effect of multiple level of two factors.
Design of experiments	Need to satisfy only two principles	All three principles needs to be satisfied

10.7. Self-ASSESSMENT QUESTIONS:

1) Explain Terminologies Related to ANOVA

2) Explain Types of ANOVA

3) Comparison between One Way and Two-Way ANOVA

10.8. SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.

2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

Dr. B. Hari Mallikarjuna Reddy

LESSON-11

ANALYSIS OF COVARIANCE

11.0. OBJECTIVES:

There are several important uses of covariance analysis in industrial and agricultural research. Some of the most important ones are:

- To control experimental error and to adjust treatment means.
- To aid in the interpretation of experimental results.
- To estimate missing data.

STRUCTURE:

11.1. INTRODUCTION:

The meaning of ANCOVA is Analysis of Covariance. It is a general linear model with one continuous outcome variable (quantitative) and one or more factor variables (qualitative). ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect on the outcome variable after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

It is well known that in designed experiments the ability to detect existing differences among treatments increases as the size of the experimental error decreases, a good experiment attempts to incorporate all possible means of minimizing the experimental error. Besides proper experimentation, a proper data analysis also helps in controlling experimental error. In situations where blocking alone may not be able to achieve adequate control of experimental error, proper choice of data analysis may help a great deal. By measuring one or more *covariates* - the characters whose functional relationships to the character of primary interest are known - the Analysis of Covariance (ANCOVA) can reduce the variability among experimental units by adjusting their values to a common value of the covariates. For example, in an animal feeding trial, the initial body weight of the animals usually differs. Using this initial body weight as a covariate, the final weights recorded after the animals have been subjected to various physiological feeds (treatments) can be adjusted to the values that would have been obtained had there been no variation in the initial body weights of the animals at the start of the experiment. Another example, in a field experiment where rodents have (partially) damaged some of the plots, covariance analysis with rodent damage as a covariate could be useful in adjusting plot yields to the levels that they should have been had there been no rodent damage in any plot.

ANCOVA requires measurement of the character of primary interest plus the measurement of one or more variables known as *covariates*. It also requires that the functional relationship of the covariates with the character of primary interest is known beforehand. Generally, a linear relationship is assumed, though other type of relationships could also be assumed.

Consider the case of a variety trial in which weed incidence is used as a covariate. With a known functional relationship between weed incidence and grain yield, the character of primary interest, the covariance analysis can adjust grain yield in each plot to a common level of weed incidence. With this adjustment, the variation in yield due to weed incidence is quantified and effectively separated from that due to varietal difference.

ANCOVA can be applied to any number of covariates and to any type of functional relationship between variables *viz.* quadratic, inverse polynomial, etc. Here we illustrate the use of covariance analysis with the help of a single covariate that is linearly related with the character of primary interest. It is expected that this simplification shall not unduly reduce the applicability of the technique, as a single covariate that is linearly related with the primary variable is adequate for most of the experimental situations in industrial and agricultural research.

11.2. CONCEPT OF ANALYSIS OF COVARIANCE:

Any scientific experiment is performed to know something that is unknown about a group of treatments and to test certain hypothesis about the corresponding treatment effect.

When variability of experimental units is small relative to the treatment differences and the experimenter do not wish to use experimental design, then just take large number of observations on each treatment effect and compute its mean. The variation around mean can be made as small as desired by taking more observations.

When there is considerable variation among observations on the same treatment and it is not possible to take an unlimited number of observations, the techniques used for reducing the variation are

i) Use of proper experimental design and

ii) Use of concomitant variables.

The use of concomitant variables is accomplished through the technique of analysis of covariance. If both the techniques fail to control the experimental variability then the number of replications of different treatments (in other words, the number of experimental units) are needed to be increased to a point where adequate control of variability is attained.

Linear model

$$Y = X_1\beta_1 + X_2\beta_2+...+X_p\beta_p +\varepsilon,$$

if the explanatory variables are quantitative variables as well as indicator variables, i.e., some of them are qualitative and some are quantitative, then the linear model is termed as analysis of covariance (ANCOVA) model.

Note that the indicator variables do not provide as much information as the quantitative variables. For example, the quantitative observations on age can be converted into indicator variable. Let an indicator variable be

$$D = \begin{cases} 1 \text{ if age} \geq 17 \text{ years} \\ 0 \text{ if age} > 17 \text{ years.} \end{cases}$$

Now the following quantitative values of age can be changed into indicator variables.

Ages (Years)	Ages
14	0
15	0
16	0
17	1
20	1
21	1
22	1

In many real applications, some variables may be quantitative and others may be qualitative. In such cases, ANCOVA provides a way out.

It helps is reducing the sum of squares due to error which in turn reflects the better model adequacy diagnostics.

See how does this work:

In one way model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, we have $TSS1_1 = SSA_1 \quad +SSE_1$

In two way model: $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, we have $TSS1_2 = SSA_2 \quad +SSE_2$

In two way model: $Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}$, we have $TSS1_3 = SSA_3 + SSB_3 + SS\gamma_3 \quad + SSE_2$

If we have a given data set, then ideally

$$TTS_1 = TSS_2 = TSS_3$$
$$SSA_1 = SSA_2 = SSA_3;$$
$$SSB_2 = SSB_3$$
$$\text{So } SSE_1 \geq SSE_2 \geq SSE_3$$

Note that in the construction of F – statistics,

$$\frac{SS(\text{effects})/df}{SSE/df}$$

So, F- statistic essentially depends on the SSEs.

Smaller SSE $\Rightarrow$ lager F $\Rightarrow$ more chance of rejection.

Since SSA, SSB etc., here are based on dummy variables, so obviously if SSA, SSB, etc. are based on quantitative variables, they will provide more information. Such ideas are used in ANCOVA models and we construct the model by incorporating the quantitative explanatory variables in ANOVA models.

In another example, suppose our interest is to compare several different kinds of feed for their ability to put weight on animals. If we use ANOVA, then we use the final weights at the end of experiment. However, final weights of the animals depend upon the initial weight of the animals at the beginning of the experiment as well as upon the difference in feeds.

Use of ANCOVA models enables us to adjust or correct these initial differences.

ANCOVA is useful for improving the precision of an experiment. Suppose response Y is linearly related to covariate X **(or concomitant variable).** Suppose experimenter cannot control X but can observe it. ANCOVA involves adjusting Y for the effect of X. If such an adjustment is not made, then the X can inflate the error mean square and makes the true differences is Y due to treatment harder to detect.

If, for a given experimental material, the use of proper experimental design cannot control the experimental variation, the use of concomitant variables (which are related to experimental material) may be effective in reducing the variability.

Consider the One-Way Classification model as

$$\begin{aligned} &E(Y_{ij} = \beta_i + \gamma t_{ij} \\ &Var(Y_{ij}) = \sigma^2 \end{aligned} \qquad i=1..... p, j=1,..., N,$$

If usual analysis of variance for testing the hypothesis of equality of treatment effects shows a highly significant difference in the treatment effects due to some factors affecting the experiment, then consider the model which takes into account this effect

$$E(Y_{ij} = \beta_i + \gamma t_{ij} + \gamma_2 t^2{}_{ij} \qquad i = 1,.........., \ j = 1,...........j$$

$$Var(Y_{ij}) = \sigma^2$$

$$E\left(Y_{ij}\right) = \mu + \alpha_i + \beta j + \gamma t_{ij} \qquad i = 1,.........., \ j = 1,...........j$$

Or

$$E\left(Y_{ij}\right) = \mu + \alpha_i + \beta j + \gamma t_{ij} + \gamma_2 w_{ij}$$

with $\sum_i \alpha_i, \sum_j \beta_j = 0,$

Where are the observations on concomitant variables (which are related to X_{ij}) and γ is the regression coefficient associated with t_{ij}. With this model, the variability of treatment effects can be considerably reduced.

For example, in any agricultural experimental, if the experimental units are plots of land then, t_{ij} can be measure of fertility characteristic of the j^th^ plot receiving i^th^ treatment and X can be yield.

In another example, if experimental units are animals and suppose the objective is to compare the growth rates of groups of animals receiving different diets. Note that the observed differences in growth rates can be attributed to diet only if all the animals are similar in some observable characteristics like weight, age etc. which influence the growth rates.

In the absence of similarity, user, which is the weight or age of j animal receiving it treatment.

If we consider the quadratic regression is given by

$$E(Y_{ij} = \beta_i + \gamma t_{ij} + \gamma_2 t^2{}_{ij} \qquad i = 1,.........., \ j = 1,...........j$$

$$Var(Y_{ij}) = \sigma^2$$

ANCOVA in this case is the same as ANCOVA with two concomitant variables and

In two-way classification with one observation per cell,

$$E\left(Y_{ij}\right) = \mu + \alpha_i + \beta j + \gamma t_{ij} \qquad i = 1,.........., \ j = 1,...........j$$

or

$$E\left(Y_{ij}\right) = \mu + \alpha_i + \beta j + \gamma t_{ij} + \gamma_2 w_{ij}$$

with $\sum_i \alpha_i, \sum_j \beta_j = 0,$

The concomitant variables can be fixed on random.

We consider the case of fixed concomitant variables only.

11.3. ONE-WAY CLASSIFICATION:

Let $Y_{ij}\left(j-1...n_i, i=1...p\right)$ be a random sample of size n_i from i^{th} normal populations with mean

$$\mu_{ij}=E\left(Y_{ij}\right)=\beta_i+\gamma t_{ij}$$

$$Var\left(Y_{ij}\right)=\sigma^2$$

Where β_i,γ and σ^2 are the unknown parameters, t_{ij} are known constants which are the observations on a concomitant variable.

The null hypothesis is

$$H_0:\beta_1=....=\beta_p.$$

Let

$$\overline{y_{ia}}=\frac{1}{n_i}\sum_j y_{ij};\overline{y}_{oj}=\frac{1}{p}\sum_i y_{ij},\overline{y}_{oo}=\frac{1}{n}\sum_i\sum_j t_{ij}$$

$$\overline{t}_{ai}=\frac{1}{n_i}\sum_j t_{ij};\overline{t}_{oj}=\frac{1}{p}\sum_i t_{ij},\overline{t}_{oo}=\frac{1}{n}\sum_i\sum_j t_{ij}$$

Under the whole parametric space $\left(\pi_\Omega\right)$, use likelihood ratio test for which we obtain the $\widehat{\beta_i}'s$ and $\hat{\gamma}$ using the least squares principle or maximum likelihood estimation as follows:

Minimize $$S=\sum_i\sum_i\left(y_{ij}-\mu_{ij}\right)^2$$

$$=\sum_i\sum_i\left(y_{ij}-\beta_i-\gamma t_{ij}\right)^2$$

$$\frac{\partial S}{\partial\beta_i}=0$$ for fixed γ

$$\Rightarrow\beta_i=\overline{y}_{io}-\gamma\overline{t_{io}}$$

Put β_i in S and minimize the function by $\frac{\partial S}{\partial\gamma}=0$.

i.e., minimize $\sum_i \sum_j \left[y_{ij} - \bar{y}_{io} - \gamma\left(t_{ij} - \bar{t}_{io}\right)\right]^2$ with respect to γ gives

$$\hat{\gamma} = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_{io}\right)\left(t_{ij} - \overline{t_{io}}\right)}{\sum_i \sum_j \left(t_{ij} - \overline{t_{io}}\right)^2}$$

Thus $\widehat{\beta_i} = \bar{y}_{io} - \hat{\gamma}\overline{t_{io}}$

$\widehat{\mu_{ij}} = \overline{\beta}_i + \hat{\gamma}\overline{t_{ij}}$

Since $y_{ij} - \hat{\mu}_{ij} = y_{ij} - \widehat{\beta_i} - \hat{\gamma} t_{ij}$

$= y_{ij} - \bar{y}_{ij} - \hat{\gamma}\left(t_{ij} - \overline{t_{io}}\right)$

We have

$$\sum_i \sum_j \left(y_{ij} - \widehat{\mu_{ij}}\right)^2 = \sum\sum\left(y_{ij} - \bar{y}_{io}\right)^2 - \frac{\left[\sum_i \sum_j \left(y_{ij} - \bar{y}_{io}\right)\left(t_{ij} - \overline{t_{io}}\right)\right]}{\sum_i \sum_j \left(t_{ij} - \overline{t_{ij}} - \overline{t_{io}}\right)^2}$$

Under $H_0 : \beta_1 = ... = \beta_p = \beta$ (say), consider $S_w = \sum_i \sum_j \left[y_{ij} - \beta - \gamma t_{ij}\right]^2$ and minimize S_w under sample space $\left(\pi_w\right)$,

$\frac{\partial S_w}{\partial \beta} = 0$,

$\frac{\partial S_w}{\partial \gamma} = 0$

$\Rightarrow \hat{\hat{\beta}} = \bar{y}_{00} - \hat{\hat{\gamma}}\overline{t_{oo}}$

$$\hat{\hat{\gamma}} = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_{oo}\right)\left(t_{ij} - \overline{t_{oo}}\right)}{\sum_i \sum_j \left(t_{ij} - \overline{t_{oo}}\right)^2}$$

$$\hat{\hat{\mu}} = \hat{\hat{\beta}} + \hat{\hat{\gamma}} t_{ij}$$

Hence

$$\sum_i \sum_j \left(y_{ij} - \hat{\hat{\mu}}_{ij} \right)^2 = \sum_i \sum_j \left(y_{ij} - \overline{y_{oo}} \right) - \frac{\left[\sum_i \sum_j \left(y_{ij} - \overline{y_{oo}} \right)\left(t_{ij} - \overline{t_{oo}} \right) \right]}{\sum_i \sum_j \left(t_{ij} - \overline{t_{oo}} \right)^2}$$

and

$$\sum_i \sum_j \left(\widehat{\mu_{ij}} - \hat{\hat{\mu}}_{ij} \right)^2 = \sum_i \sum_j \left[\left(\overline{y}_i - \overline{y_{oo}} \right) + \hat{\gamma}\left(t_{ij} - \overline{t_{io}} \right) - \hat{\hat{\gamma}}\left(t_{ij} - \overline{t_{oo}} \right) \right]^2$$

The likelihood ratio test statistic in this case is given by

$$\lambda = \frac{\max\limits_{w} L\left(\beta, \gamma, \sigma^2\right)}{\max\limits_{\Omega} L\left(\beta, \gamma, \sigma^2\right)}$$

$$= \frac{\sum_i \sum_j \left(\hat{\mu}_{ij} - \hat{\hat{\mu}}_{ij} \right)^2}{\sum_i \sum_j \left(y_{ij} - \hat{\mu}_{ij} \right)^2}$$

Now we use the following theorems:

Theorem 1:

Let $Y = \left(Y_1, Y_2, ..., Y_n\right)'$ follow a multivariate normal distribution $N\left(\mu, \Sigma\right)$ with mean vector μ and positive definite covariance matrix Σ. Then Y'AY follows a noncentral chi-square distribution with p degrees of freedom and non-centrality parameter $\mu' A \mu$, i.e., $x^2\left(p, \mu' A \mu\right)$ if and only if ΣA is an idempotent matrix of rank p.

Theorem 2:

Let $Y = \left(Y_1, Y_2, ..., Y_n\right)'$ follows a multivariate normal distribution $N\left(\mu, \Sigma\right)$ with mean vector μ and positive definite covariance matrix Σ. Let $Y' A_1 Y$ follows $x^2\left(p_1, \mu' A_1 \mu\right)$ and $Y' A_2 Y$ follows $x^2\left(p_2, \mu' A_2 \mu\right)$. Then $Y' A_1 Y$ and $Y' A_2 Y$ are independently distributed if $A_1 \Sigma A_2 = 0$.

Theorem 3:

Let $Y = \left(Y_1, Y_2, ..., Y_n\right)'$ follows a multivariate normal distribution $N\left(\mu, \sigma^2 I\right)$, then the maximum likelihood (or least squares) estimator $L'\hat{\beta}$ of estimable linear parametric function is independently distributed of $\hat{\sigma}^2 ; L\hat{\beta}$ follow $N\left[L'\beta,\, L'\left(X'X\right)^{-1} L\right]$ and $\frac{n\hat{\sigma}^2}{\sigma^2}$ follows $x^2\left(n-p\right)$where rank (X) = p.

Using these theorems on the independence of quadratic forms and dividing the numerator and denominator by respective degrees of freedom, we have

$$F = \frac{n-p-1 \sum\limits_i \sum\limits_j \left(\hat{\mu}_{ij} - \hat{\hat{\mu}}_{ij}\right)^2}{p-1 \sum\sum \left(y_{ij} - \hat{\mu}_{ij}\right)^2} \sim F\left(p-1,\, n-p\right) \text{ under } H_0$$

So, reject H_0 whenever $F \geq F_{1-\alpha}\left(p-1,\, n-p\right)$ at α level of significance.

The terms involved in λ can be simplified for computational convenience follows:

We can write

$$\sum_i \sum_j \left(y_{ij} - \hat{\hat{\mu}}_{ij}\right)^2$$

$$= \sum_i \sum_j \left[y_{ij} - \hat{\hat{\beta}} - \hat{\hat{\gamma}} t_{ij}\right]^2$$

$$= \sum_i \sum_j \left[\left(y_{ij} - \overline{y}_{oo}\right) - \hat{\hat{\gamma}}\left(t_{ij} - \overline{t_{oo}}\right)\right]^2$$

$$= \sum_i \sum_j \left[\left(y_{ij} - \overline{y}_{oo}\right) - \hat{\hat{\gamma}}\left(t_{ij} - \overline{t_{oo}}\right) + \hat{\gamma}\left(t_{ij} - \overline{t_{oo}}\right) - \hat{\gamma}\left(t_{ij} - \hat{t}_{io}\right)\right]^2$$

$$= \sum_i \sum_j \left[\left(y_{ij} - \overline{y}_{io}\right) - \hat{\gamma}\left(t_{ij} - \overline{t_{io}}\right)\right]^2$$

$$= \sum_i \sum_j \left[\left(y_{ij} - \overline{y}_{io}\right) + \hat{\gamma}\left(t_{ij} - \overline{t_{io}}\right) - \hat{\hat{\gamma}}\left(t_{ij} - \overline{t_{oo}}\right)\right]^2$$

$$= \sum_i \sum_j \left(y_{ij} - \hat{\mu}_{ij} \right)^2 + \sum_i \sum_j \left(\hat{\mu}_{ij} - \widehat{\hat{\mu}}_{ij} \right)$$

For Computational Convenience

$$\lambda = \frac{\sum_i \sum_j \left(\widehat{\mu}_{ij} - \widehat{\widehat{\mu}}_{ij} \right)^2}{\sum_i \sum_j \left(y_{ij} - \widehat{\mu}_{ij} \right)^2} = \frac{\left(T_{yy} - \frac{T_{yt}^2}{T_{it}} \right) - \left(E_{yy} - \frac{E_{yt}^2}{E_{tt}} \right)}{\left(E_{yy} - \frac{E_{yt}^2}{E_{yy}} \right)}$$

Where

$$T_{yy} = \sum_i \sum_j \left(y_{ij} - \overline{y_{oo}} \right)^2$$

$$T_{tt} = \sum_i \sum_j \left(t_{ij} - \overline{t_{oo}} \right)^2$$

$$T_{yt} = \sum_i \sum_j \left(y_{ij} - \overline{y_{oo}} \right)\left(t_{ij} - \overline{t_{oo}} \right)$$

$$E_{yy} = \sum_i \sum_j \left(y_{ij} - \overline{y_{io}} \right)^2$$

$$E_{tt} = \sum_i \sum_j \left(t_{ij} - \overline{t_{io}} \right)^2$$

$$E_{yt} = \sum_i \sum_j \left(y_{ij} - \overline{y_{io}} \right)\left(t_{ij} - \overline{t_{io}} \right)$$

Analysis of Covariance Table for One-Way Classification is as follows:

Source of Variation	**Degrees of Freedom**	**Sum of Products** **yyyt tt**	**Adjusted Sum of Squares**		**F**
			Degrees of Freedom	**Sum of Squares**	
Population	p-1	$P_{yy}\left(= T_{yy} - E_{yy}\right) P_{yt}$ $\left(= T_{yt} - E_{yt}\right) P_{it}\left(= T_{it} - E_{it}\right)$	P-1	$q_1 = q_0 - q_2$	$\frac{n-p-1}{p-1} \frac{q_1}{q_2}$
Error	n-p	$E_{yy}\ E_{yt}\ E_{tt}$	n-p-1	$q_2 = E_{yy} - \frac{E_{yt}^2}{E_{yy}}$	
Total	n-1	$T_{yy}\ T_{yt}\ T_{tt}$	n-2	$q_0 = T_{yy} - \frac{T_{yt}^2}{T_{tt}}$	

If H_0 is rejected, employ multiple comprises methods to determine which of the contrasts in β_i are responsible for this.

For any estimable linear parametric contrast

$$\varphi = \sum_{i=1}^{p} C_i\beta_i \text{ with} \sum_{i=1}^{p} C_i = 0,$$

$$\hat{\varphi} = \sum_{i=1}^{p} C_i\hat{\beta}_i = \sum_{i=1}^{p} C_i\overline{y}_i - \hat{\gamma}\sum_{i=1}^{p} C_{ii}\overline{t}_i$$

$$\text{Var}\left(\hat{\gamma}\right) = \frac{\sigma^2}{\sum_i\sum_j\left(t_{ij} - \overline{t}_i\right)^2}$$

$$\Rightarrow \text{Var}\left(\hat{\varphi}\right) = \sigma^2\left[\sum_i \frac{C_i^2}{n_i} + \frac{\left(\sum_i C_i\overline{t}_i\right)^2}{\sum_i\sum_j\left(t_{ij} - \overline{t}_i\right)^2}\right]$$

11.4. TWO-WAY CLASSIFICATION (WITH ONE OBSERVATIONS PER CELL):

Consider the case of two-way classification with one observation per cell.

Let $y_{ij} \sim N\left(\mu_{ij}, \sigma^2\right)$ be independently distributed with

$$E\left(y_{ij}\right) = \mu + \alpha_i + \beta_i + \gamma t_{ij},\ i = 1...I,\ j = 1...J$$

$$V\left(y_{ij}\right) = \sigma^2$$

Where

μ: Grand mean

α_1 : Effect of i^{th} level of A satisfying $\sum_i^{I} \alpha_i = 0$

β_1 : Effect of j^{th} level of B satisfying $\sum_i^{J} \beta_j = 0$

t_{ij} : observation (known) on concomitant variable.

The null hypothesis under consideration are

$$H_{0\alpha} : \alpha_1 = \alpha_2 = ... = \alpha_I = 0$$

$H_{0\beta} : \beta_1 = \beta_2 = \ldots = \beta_J = 0$

Dimension of whole parametric space $\left(\pi_\Omega\right) : I + J$

Dimension of sample space $\left(\pi_{\omega\alpha}\right) : J + 1$ under $H_{0\alpha}$

Dimension of sample space $\left(\pi_{w\beta}\right) : I + 1$ under $H_{0\beta}$

With respective alternative hypotheses as

$H_{1\alpha}$: At least one pair of α's is not equal

$H_{1\beta}$: At least one pair of β's is not equal.

Consider the estimation of parameters under the whole parametric space $\left(\pi_\Omega\right)$

Find minimum value of $\sum_i \sum_j \left(y_{ij} - \mu_{ij}\right)^2$ under π_Ω

To do this, Minimize

$$\sum_i \sum_j \left(y_{ij} - \mu - \alpha_i - \beta_j - \gamma t_{ij}\right)^2$$

For fixed γ, which gives on solving the least squares estimates (or the maximum likelihood estimates) of the respective parameters as

$$\mu = \overline{y_{oo}} - \gamma \overline{t_o}$$

$$\alpha_i = \overline{y_i} - \overline{y_{oo}} - \gamma\left(\overline{t_{oo}} - \overline{t_{oo}}\right) \qquad \ldots. (1)$$

$$\beta_j = \overline{y_{oj}} - \overline{y_{oo}} - \gamma\left(\overline{t_{oj}} - \overline{t_{oo}}\right)$$

Under these values of μ, α_i and β_j, the sum of squares $\sum_i \sum_j \left(y_{ij} - \mu - \alpha_i - \beta_j - \gamma t_{ij}\right)^2$ reduces to

$$\sum_i \sum_j \left[y_{ij} - \overline{y_{oo}} - \overline{y_{oj}} + \gamma\left(t_{ij} - \overline{t_{io}} - \overline{t_{oj}} + \overline{t_{oo}}\right)\right]^2 \quad \ldots..(2)$$

Now minimization of (2) with respect to γ gives

$$\hat{\gamma} = \frac{\sum_{i=1}\sum_{j=1}\left(y_{ij} - \overline{y_{io}} - \overline{y_{oj}} + \overline{y_{oo}}\right)\left(t_{ij} - \overline{t_{oj}} + \overline{t_{oo}}\right)}{\sum_{i=1}\sum_{j=1}\left(t_{ij} - \overline{t_{io}} - \overline{t_{oj}} + \overline{t_{oo}}\right)^2}$$

Using $\hat{\gamma}$, we get from (1)

$$\hat{\mu} = \overline{y_{oo}} - \hat{\gamma}\bar{t}_{oo}$$

$$\hat{\alpha}_i = \left(\bar{y}_{oo} - \bar{y}_{oo}\right) - \hat{\gamma}\left(\bar{t}_{oo} - \bar{t}_{oo}\right)$$

$$\hat{\beta}_j = \left(\bar{y}_{oj} - \bar{y}_{oo}\right) - \hat{\gamma}\left(\bar{t}_{oj} - \bar{t}_{oo}\right)$$

Hence

$$\sum_i\sum_j\left(y_{ij} - \hat{\mu}_{ij}\right)^2$$

$$= \sum_i\sum_j\left(y_{ij} - \bar{y}_{io} - \bar{y}_{oo}\right) - \frac{\left[\sum_i\sum_j\left(y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo}\right)\left(t_{ij} - \bar{t}_{oj} + \bar{t}_{oo}\right)\right]}{\sum_i\sum_j\left(t_{ij} - \bar{t}_{io} - \bar{t}_{oj} + \bar{t}_{oo}\right)^2}$$

$$= E_{yy} - \frac{E_{yt}^2}{E_{tt}}$$

Where

$$E_{yy} = \sum_i\sum_j\left(y_{ij} - y_{io} - \bar{y}_{oj} + \bar{y}_{oo}\right)^2$$

$$E_{yt} = \sum_i\sum_j\left(y_{ij} - y_{io} - \bar{y}_{oj} + \bar{y}_{oo}\right)\left(t_{ij} - \bar{t}_{io} - \bar{t}_{oj} + \bar{t}_{oo}\right)$$

$$E_{tt} = \sum_i\sum_j\left(t_{ij} - \bar{t}_{io} - \bar{t}_{oj} + \bar{t}_{oo}\right)^2$$

Case (i): Test of $H_{0\alpha}$

Minimize $\sum_i\sum_j\left(y_{ij} - \mu - \beta_j - \gamma t_{ij}\right)^2$ with respect to μ, β_j and γ gives the least squares estimates) (or the maximum likelihood estimates) of respective parameters as

$$\Rightarrow \hat{\hat{\mu}} = \bar{y}_{oo} - \hat{\hat{\gamma}}\overline{t_{oo}}$$

$$\hat{\hat{\beta}}_j = \bar{y}_{oj} - \bar{y}_{oo} - \hat{\hat{\gamma}}\left(\bar{t}_{oj} - \bar{t}_{oo}\right)$$

$$\hat{\hat{\gamma}} = \frac{\sum_i \sum_j \left(y_{ij} - \bar{y}_{oj}\right)\left(t_{ij} - \bar{t}_{oj}\right)}{\sum_i \sum_j \left(t_{ij} - \bar{t}_{oj}\right)^2} \qquad \text{..... (3)}$$

$$\hat{\hat{\mu}} = \hat{\hat{\mu}} + \hat{\hat{\beta}}_j + \hat{\hat{\gamma}} t_{ij}$$

Substituting these estimates in (3) we get

$$\sum_i \sum_j \left(y_{ij} - \hat{\hat{\mu}}_g\right)^2 = \sum_i \sum_j \left(y_{ij} - \bar{y}_j\right)^2 - \frac{\left[\sum_i \sum_j \left(y_{ij} - \bar{y}_{oj}\right)\left(t_{ij} - \bar{t}_{oj}\right)\right]}{\sum_i \sum_j \left(t_{ij} - \bar{t}_{oj}\right)^2}$$

$$= E_{xy} + A_{yy} - \frac{\left[E_{yt} + A_{yt}\right]}{E_{tt} + A_{tt}}$$

Where

$$A_{yy} = \sum_i J\left(\bar{y}_{io} - \bar{y}_{oo}\right)^2$$

$$A_{tt} = \sum_i J\left(\bar{t}_{io} - \bar{t}_{oo}\right)^2$$

$$A_{yt} = \sum_i J\left(\bar{y}_{io} - \bar{y}_{oo}\right)\left(\bar{t}_{io} - \bar{t}_{oo}\right)^2$$

$$E_{yy} = \sum_i \sum_j \left(y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo}\right)^2$$

$$E_{tt} = \sum_i \sum_j \left(t_{ij} - \bar{t}_{io} - \bar{t}_{oj} + \bar{t}_{oo}\right)^2$$

$$E_{yt} = \sum_i \sum_j \left(y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo}\right)\left(t_{ij} - \bar{t}_{io} - \bar{t}_{oj} + \bar{t}_{oo}\right)$$

Thus, the likelihood ratio test statistic for testing $H_{0\alpha}$ is

$$\lambda_1 = \frac{\sum_i \sum_j \left(y_{ij} - \hat{\hat{\mu}}_{ij}\right)^2 - \sum_i \sum_j \left(y_{ij} - \hat{\mu}_{ij}\right)^2}{\sum_i \sum_j \left(y_{ij} - \hat{\mu}_{ij}\right)^2}$$

Adjusting with degrees of freedom and using the earlier result for the independence of two quadratic forms and their distribution

$$F_1 = \frac{(IJ-I-J)}{(I-1)}\left[\frac{\sum_i\sum_j\left(y_{ij}-\hat{\hat{\mu}}_{ij}\right)^2 - \sum_i\sum_j\left(y_{ij}-\hat{\mu}_{ij}\right)^2}{\sum_i\sum_j\left(y_{ij}-\hat{\mu}_{ij}\right)^2}\right] \sim F(I-1,\ IJ-I-J) \text{ under}$$

$H_{o\alpha}$

So, the decision rule is to reject $H_{a\alpha}$ whenever $F_1 > F_{1-\alpha}(I-1,\ IJ-I-J)$.

Case (ii): Test of $H_{o\beta}$

Minimize $\sum_i\sum_j\left(y_{ij}-\mu-\alpha_i-\gamma t_{ij}\right)^2$ with respect to $\mu,\ \alpha_i$ and γ gives the least squares estimates (or maximum likelihood estimates) of respective parameters as

$$\hat{\mu} = \bar{y}_{oo} - \hat{\gamma}\bar{t}_{oo}$$

$$\hat{\alpha}_j = \bar{y}_{io} - \bar{y}_{oo} - \hat{\gamma}\left(\bar{t}_{io} - \bar{t}_{oo}\right)$$

$$\hat{\gamma} = \frac{\sum_i\sum_j\left(y_{ij}-\bar{y}_{io}\right)\left(t_{ij}-\bar{t}_{io}\right)}{\sum_i\sum_j\left(t_{ij}-\bar{t}_{io}\right)^2} \qquad \text{….. (4)}$$

$$\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_{ij}$$

From (4), we get

$$\sum_i\sum_j\left(y_{ij}-\hat{\mu}_{ij}\right)^2 = \sum_i\sum_j\left(y_{ij}-\bar{y}_{io}\right)^2 - \frac{\left[\sum_i\sum_j\left(y_{ij}-\bar{y}_{io}\right)\left(t_{ij}-\bar{t}_{oj}\right)\right]}{\sum_i\sum_j\left(t_{ij}-\bar{t}_{io}\right)^2}$$

$$= E_{yy} + B_{yy} - \frac{\left[E_{jt}+B_{yt}\right]^2}{B_{tt}}$$

$$B_{yy} = \sum_{j} I\left(\bar{y}_{oj} - \bar{y}_{oo}\right)^2$$

Where $B_{tt} = \sum_{j} I\left(\bar{t}_{oj} - \bar{t}_{oo}\right)^2$

$$B_{yt} = \sum_{j} I\left(\bar{y}_{io} - \bar{y}_{oo}\right)\left(\bar{t}_{oj} - \bar{t}_{oo}\right)^2$$

Thus, the likelihood ratio test statistic for testing $H_{0\beta}$ is

$$F_2 = \frac{(IJ-I-J)}{(J-1)}\left[\frac{\sum_{i}\sum_{j}\left(y_{ij} - \bar{\mu}_{ij}\right)^2 - \sum_{i}\sum_{j}\left(y_{ij} - \hat{\mu}_{ij}\right)^2}{\sum_{i}\sum_{j}\left(y_{ij} - \hat{\mu}_{ij}\right)^2}\right] \sim F(J-1,\ IJ-I-J) \text{ under}$$

$H_{o\beta}$.

So, the decision rule is to reject $H_{0\beta}$ whenever $F_2 \geq F_{1-\alpha}\left(Y-1,\ IJ-I-J\right)$

If $H_{o\alpha}$ is rejected, use multiple comparison methods to determine which of the contrasts α_i are responsible for this rejection. The same is true for $H_{o\beta}$.

The Analysis of Covariance Table for Two-Way Classification is as follows:

Source of Variation	**Degrees of Freedom**	**Sum of Products**				**F**
		yy	**yt**	**tt**		
Between levels of A	I-1	A_{yy}	A_{yt}	A_{tt}	I-1 q_0=q_3-q	$F_1 = \frac{IJ-I-J}{I-1}\frac{q_o}{q_2}$
Between levels of B	J-1	B_{yy}	B_{yt}	B_{tt}	J-1 q_1=q_4-q_2	$F_2 = \frac{IJ-I-J}{J-1}\frac{q_1}{q_2}$
Error	(I-1) (J-1)	E_{yy}	Y_{et}	E_{tt}	IJ-I-J $q_2 = E_{yy} - \frac{E_{yt}^2}{E_{tt}}$	
Total	IJ-1	T_{yy}	T_{yt}	T_{tt}	IJ-2	
Error + levels of A	IJ-J				$q_3 = \left(A_{yy} + E_{yy}\right) - \frac{\left(A_{yt} + E_{yt}\right)^2}{A_{tt} + E_{tt}}$	
Error + levels of B	IJ-I				$q_4 = \left(B_{yy} + E_{yt}\right) - \frac{\left(B_{yt} + E_{yt}\right)^2}{B_{tt} + E_{tt}}$	

11.5. SELF-ASSESSMENT QUESTIONS:

1) Explain difference between ANOVA and ANCOVA
2) Explain ANCOVA one-way classification with one observation per cell
3) Explain ANCOVA two-way classification with one observation per cell
4) Explain analysis of covariance with a single concomitant variable.

11.6. SUGGESTED READINGS:

1) Kempthorne, O, (1951), *The design and Analysis of Experiments*, Wiley Eastern Private Limited**.**
2) Federer, Wt (1967), *Experimental Design Theory and Application*, Oxford & IBH Publishing Company.
3) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
4) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.
5) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

Dr. B. Hari Mallikarjuna Reddy

LESSON-12

COMPLETELY RANDOMISED DESIGN

12.0. OBJECTIVES:

After studying this unit, you would be able to

- Describe the experimental design;
- Explain the planning and classification of experimental designs;
- Describe the principles of design of experiments;
- Explain the completely randomized design;
- Describe the layout of CRD;
- Explain the statistical analysis of CRD; and
- Explain the advantages and disadvantages as well as the suitability of CRD.

STRUCTURE:

12.1. INTRODUCTION:

The modern concepts of experimental designs were primarily given by Ronald A. Fisher in the 1920s and 1930s at “Rothamasted Experimental Station”, an agricultural research station of London. In Fisher’s first book on design of experiments, he showed how valid conclusions could be drawn efficiently from experiments with natural fluctuation such as temperature, soil conditions and rainfall, that is, in the presence of nuisance variables. The known nuisance variables usually cause systematic biases in groups of results (e.g. batch-to-batch variables). The unknown nuisance variables usually cause random variability in the results and are called inherent variability or noise.

The experimental design was first used in an agricultural context, the method has been applied successfully in the military and in industry since the 1940s. Besse Day, working at U. S. Naval Experimentation Laboratory, used experimental designs to solve problems such as finding the cause of bad welds at the naval shipyards during World War II. George Box, employed by Imperial Chemical Industries before coming to the United States, is a leading developer of experimental design produced for optimizing chemical process. W. Edwards Deming taught statistical methods, including experimental designs, to Japanese scientist and engineers in the early 1950's at a time when "Made in Japan" meant poor quality. Ganache Taguchi, the most well-known of this group of Japanese scientists is famous for his quality improvement methods. One of the companies where Taguchi first applied his methods was Toyota. Since the late 1970's, U.S. industry has become interested again inquality improvement initiatives, now known as "Total Quality" & "Six-sigma" programs. Design of experiments is considered an advanced method in the six sigma programs, which were pioneered at Motorola & GE.

According to Bernad Ostle, "The design of experiment is, the complete sequence of steps taken ahead of time to ensure that the appropriate data will be obtained in a way which permits an objective analysis to valid inferences with respect to stated problem".

In any field of study either in life sciences or some other, it is essential to plan an experiment, i.e. what is the object and which type of data is required. In order to make use of time and energy spent on experiment, it should be planned with a careful designing. Once a design of experiment is decided, the observations are obtained from it and with the technique of analysis of variance, the data is analyzed.

12.1.1. Planning of an Experiment:

There are some basic points regarding the planning of an experiment, which should be under consideration. These are as follows:

1. The Experiment should be Free from Bias:

An experiment must be planned so that it gives an unbiased estimate of the values we wish to measure. It is a matter of the design being such that no bias on the part of the experimenter can possibly enter into the results. This is achieved mainly by randomisation.

2. There must be a Measure of Error:

The true experiment is one that is strictly objective. It should furnish a measure of error and this error alone should be the measuring stick of significance.

3. There must be a Clearly Defined Objective:

For an experiment it is essential to specify the objects perfectly. In other words, the objective of the experiment should be clearly defined.

4. The Experiment should have Sufficient Accuracy:

The accuracy of an experiment can be brought by the elimination of technical errors and by increasing replications. The number of replications should be decided to produce a given degree of accuracy.

12.1.2. Classification of Experimental Designs:

Statisticians by themselves do not design experiments, but they have developed a number of structured schedules called “experimental designs”, which they recommend for the taking of measurements. These designs have certain rational relationships to the purposes, needs and physical limitations of experiments. Designs also offer certain advantages in economy of experimentation and provide straightforward estimates of experimental effects and valid estimates of variance. There are a number of ways in which experiment designs might be classified, for example, the following:

1) By the number of experimental factors to be investigated (e.g., single- factor versus multifactor designs)

2) By the structure of the experimental design (e.g., blocked, factorial, nested, or response-surface design)

3) By the kind of information which the experiment is primarily intended to provide (e.g. estimates of effects, estimates of variance, or empirical mappings).

12.2. BASIC DEFINITIONS OF EXPERIMENTAL DESIGN:

Several fundamental terms are widely used throughout this section. They may be defined as follows:

1. Treatment:

In an experiment, there are some variants under study, the effects of which are measured and tested (compared). These variants will be referred to as treatments. For example, to test the effects of three fertilizers, i.e., Nitrogen, Phosphorus and Potash on the yield of a certain crop. Then Nitrogen, Phosphorus and Potash are called treatments.

2. Yield:

The response of the treatment is measured by some indicator such as crop production, milk production, body temperature, mileage of engine set, etc. Such an indicator is called yield. The treatments are applied to some units such as field plots, sample of cows, sample of patients, sample of engine, sets, etc. and the effect on the yield is observed.

3. Experimental Units:

A unit to which one treatment applied is called experimental unit. It is the smallest division of an experimental material to which the treatment applied and on which the variable under study is measured. In carrying out an experiment, we should clear as to what constitute the experimental unit.

It can be understood that in a field of agriculture it is called plot, in the field of animal husbandry it may be a cow (cattle), in the field of medicine it may be a patient and in the field of automobile industry it may be engine set and so on.

4. Experimental Material:

We have already explained the concept of experimental unit. The experimental material is nothing but a set of experimental units. For example, a piece of land, a group of cows, a number of patients and a group of engine sets, etc. Actually, an experimental material is that material on which some set of treatments are applied and tested.

5. Blocks:

The experimental material is divided into a number of groups or strata which are so formed that they are within homogeneous and between heterogeneous. These groups or strata are called blocks.

6. Experimental Error:

There is always a variation between the yields of the different plots even when they get the same treatment. This variation exists due to non- assignable causes, which cannot be detected and explained. These are taken to be of random type. This unexplained random part of variation is termed as experimental error. This include all types of extraneous variation due to, (i) inherent variability in the experimental units, (ii) error associated with the measurement made and (iii) lack of representativeness of the sample of the population understudy.

7. Precision:

The precision of an experiment is measured by the reciprocal of the variance of a mean, i.e.

$$\frac{1}{v(\bar{x})} = \frac{1}{\sigma_{\bar{x}}^2} = \frac{n}{\sigma^2}$$

As n, the replication number increases, precision also increases.

8. Uniformity Trial:

We know that to increase the efficiency of a design, the plots should be arranged into homogeneous blocks. It can be done only if we have a correct idea about the fertility variation of the field. This is achieved through uniformity trial. It is known that fertility of soil does not increase or decrease uniformly in any direction but it is distributed over the entire field in an erratic manner. By a uniformity trial, we mean a trial in which the field (experimental material) is divided into small units (plots) and the same treatment is applied on each of the units and their yields are recorded. From these yields we can draw a fertility control map which gives us a graphic picture of the variation of the soil fertility and enables us to form a good idea about the nature of the soil fertility variation. This fertility control map is obtained by joining the points of equal fertility through lines.

A Uniformity Trial gives us an idea about the

1) Fertility gradient of the field,
2) Determination of the shape of the plots to be used,
3) Optimum size of plots,
4) Estimation of number of replications required for achieving certain degree of accuracy.

12.3. PRINCIPLES OF DESIGN OF EXPERIMENTS:

Good experimentation is an art and depends heavily upon the prior knowledge and abilities of the experimenter. Designing an experiment means deciding how the observations or measurements should be taken to answer a particular question in a valid, efficient and economical way. If a design is properly designed, then there will exists an appropriate way of anal sing the data. From an ill-designed experiment, no conclusion can be drawn.

The fundamental principles in design of experiments are the solutions to the problems in experimentation posed by the two types of nuisance factors and serve to improve the efficiency of experiments. For the validity of the design Prof. R.A. Fisher gave three principles of design of experiments, those fundamental principles are:

- Randomization
- Replication
- Local Control

12.3.1. Randomization:

The principle of randomization is essential for a valid estimate of the experimental error and to minimize the bias in the results. In the words of Cochran and Cox, “Randomization is analogous to insurance in that it is a precaution against disturbances that may or may not occur and they may or may not be serious if they do occur”. Thus, randomization is so done that each treatment should get an equal chance. We mean that the treatments should be allocated randomly, i.e., by the help of random numbers. The following are the advantages of randomizations:

1) It provides a basis for the test of significance because randomization ensures the independence of the observations which is one of the assumptions for the analysis of variance.
2) It is also a device for eliminating bias. Bias creeps in experiment, when the treatments are not assigned randomly to the units. This bias may be personal or subjective. The randomization ensures the validity of the results.

12.3.2. Replication:

“Replication” is the repetition, the rerunning of an experiment or measurement in order to increase precision or to provide the means for measuring precision. A single replicate consists of a single observation or experimental run. Replication provides an opportunity for the effects of uncontrolled factors or factors unknown to the experimenter to balance out and thus, through randomization, acts as a bias-decreasing tool. Suppose a pain-relieving drug A is applied to 4 patients, we say that drug A is replicated four times. By repeating a treatment, it is possible to obtain a more reliable estimate because it reduces the experimental error. Further by repeating a treatment number of times we can judge the average performance of a treatment and the situation becomes clearer. Basically, there are following uses of replication:

1) It enables us to obtain a more precise estimate of the treatment's effects.

2) The next important purpose of replication is to provide an estimate of the experimental error without which we cannot test the significance of the difference between any two treatments. The estimate of experimental error is obtained by considering the difference in the plots receiving the same treatment in different replications and there is no other alternative of obtaining this estimate.

3) For a desired amount of precision, the minimum number of replications can be obtained.

12.3.3. Local Control:

This method is used to attain the accuracy or to reduce the experimental error without increasing unduly the number of replications. Local control is a technique that handles the experimental material in such a way that the effects of variability are reduced. In local control, experimental units are divided into a number of homogeneous groups called blocks. These blocks are so formed that they are homogeneous within and heterogeneous between. This blocking of experiment may be row-wise, column-wise or both according to the number of factors responsible for heterogeneity. Different types of blocking constitute different types of experimental designs. The following are the advantages of local control:

1) By means of local control, the experimental error is reduced considerably and the efficiency of the design is increased.

2) By means of local control the test procedure becomes more sensitive or powerful.

Besides the above three principles, there are some other general principles in designing an experiment. Familiarity with the treatments and experimental material is an asset. Selection of experimental site is an asset. Selection of experimental site should be carefully done. Within block variability should be reduced.

12.4. SIZE AND SHAPE OF THE PLOTS:

In field experiments, the size and shape of plots as well as of blocks influence the experimental error. The total available experimental area remaining fixed, an increase in size of plots will automatically decrease the number of plots and indirectly increases the block size. In order to reduce the flow of experimental material from one plot to another, it is customary to leave out strips of land between consecutive plots and also between blocks. These non- experimental areas are known as guard area. The size and shape of the plot should be such that we make a compromise between statistical and practical requirements i.e. if plot size is x and the variance of the plot is V(x), then V(x) is minimum (statistical consideration) and there should be no disturbance for agricultural operations (practical requirements).

The size and shape of block will ordinarily be determined by the size and shape of plots and the number of plots in a block. It is desirable from the point of view of error control to have small variations among the plots within a block and large variation among the blocks i.e. in general the division of experimental material into blocks is made in such a way that plots within blocks are as homogeneous as possible.

Different Experimental Designs:

Based on these fundamental principles, we have certain designs. The analysis of those designs is based on the theory of least squares which gives the best estimates of the treatments effects and was initiated by Fisher (1926) followed by Yates (1936), Bose & Nair (1939) and Rao (1976). The following three designs are frequently used:

1) Completely Randomized Design

2) Randomized Block Design

3) Latin Square Design

12.5. COMPLETELY RANDOMISED DESIGN:

The simplest of all the design is completely randomized design (CRD) which is applied in the case when the experimental materials are homogeneous. CRD is based on two principles i.e. randomization and replication. The third principle, i.e. local control is not used because it is assumed that experimental materials are homogeneous. In this, the treatments are allocated randomly to the experimental units and each treatment is assigned to different experimental units completely at random (can be repeated any number of times) that is why it is called completely randomized design.

Suppose we have k treatments under comparison and the ith treatment is to be replicated n_i times for i =1, 2, …, k, then the total number of units required for the design are $n = \sum_{i=1}^{k} n_i$.

We allocate the k treatments completely at random to n_i units such that ith treatment appears n_i times in the experiment.

12.5.1. Layout of Completely Randomized Design:

The term layout refers to the allocation of different treatment to the experimental units. We have already said that treatments are allocated completely at random to the different experimental units. Every experimental unit has the same chance of receiving a particular treatment.

Suppose we want to test the effect of three pain relieving drugs A, B and C on twelve patients. Then we first number all the patients (units) from 1 to 12. Then from a random number table of one digit we pick up 12 numbers which are less than 4. Suppose the numbers are 1, 3, 2, 1, 3, 2, 1, 3, 2, 2, 3, 1. Thus the drug A is allotted to patient 1, drug C is allotted to patient number 2 and so on. It can be shown below:

(1) A	(2) C	(3) B	(4) A	(5) C	(6) B
(1) A	(2) C	(3) B	(4) B	(5) C	(6) A

It is clear from the above layout that the replications of A, B and C are equal. If the number of replications for each treatment is 5, 4 and 3 respectively, we number the experimental units in a convenient way from 1 to 12. We then get a random permutation of the experimental units. To the first 5 of the units in the random permutation we assign treatment A, to the next 4 units' treatment B is assigned and the treatment C is assigned to the remaining 3 units.

12.5.2. Statistical Analysis of Completely Randomized Design:

Statistical analysis of a CRD is analogous to the ANOVA for one-way classified data for fixed effect model, the linear model (assuming various effect to be additive) becomes

$$y_{ij} = \mu + \alpha_i + e_{ij},\ i = 1, 2, 3, \ldots.., k;\ j = 1, 2, 3, \ldots, ni \qquad \ldots (1)$$

where y_{ij} is the yield or response from the jth unit receiving the ith treatment, $\mu$is the general mean effect, α_i is the effect due to the ith treatment, where μ andα_i are constants so that $\sum_{i=1}^{k} n_i\alpha_i = 0$and e_{ij} is identically and independently distributed (i.i.d.) $N\left(0,\ \alpha_e^2\right)$. Then, $n = \sum_{i=1}^{k} n_1$ is the total number of experimental units.

The analysis of model given in equation (1) is as same as that of fixed effect model of one-way classified data, discussed in Unit 6 of MST-005. If we write

$$\sum_i \sum_j y_{ij} = y_{..} = G \text{ = Grand total of the n observations, and}$$

$$\sum_{j=1}^{n_i} y_{ij} = y_i = T_i \text{ = Total response in the units receiving the i}^{th} \text{ treatment,}$$

Then, as in ANOVA (one-way classified data),

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2+\sum_{i=1}^{k}n_i\left(\bar{y}_i-\bar{y}_{..}\right)^2$$

i.e. TSS = SSE + SST

where, TSS, SST and SSE are the Total Sum of Squares, Sum of Squares due to Treatments (between treatments SS) and Sum of Square due to Error (within treatment SS) given respectively by

$$TSS=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2$$

$$SST=\sum_{i=1}^{k}n_i\left(\bar{y}_i-\bar{y}_{..}\right)^2=S_T^2$$

$$\text{and } SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2=S_E^2$$

ANOVA TABLE FOR CRD

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ratio (F)**
Treatments	k-1	$SST=S_T^2$	$MSST=\frac{S_T^2}{(k-1)}$	$F_T=\frac{MSST}{MSSE}$
Error	n-k	$SSE=S_E^2$	$MSSE=\frac{S_E^2}{(n-k)}$	
Total	n-1	$S_T^2+S_E^2$		

Under the null hypothesis, $H_0:\alpha_1=\alpha_2=....=\alpha_k$ against the alternative that all α's is not equal, the test statistic

$$F_T=\frac{MSST}{MSSE}\sim f\left(k-1,\, n-k\right)$$

i.e., FT follows F distribution with (k-1, n- k) df.

If $F_T > F_{(k-1,\, n-k)}(\alpha)$ then H0 is rejected at α level of significance and we conclude that treatments differ significantly. If $F_T > F_{(k-1,\, n-k)}(\alpha)$ then H₀ may be accepted i.e. the data do not provide any evidence to prefer one treatment to the other and as such all of them can be considered alike.

If the treatments show significant effect, then we would be interested to find out which pair of treatments differs significantly. For this instead of calculating Student's t-test for different pairs of treatment means we calculate the least significant difference at the given level of significance. This least difference is called as critical different (CD) and CD at α level of significance is given by

CD = Standard error of difference between two treatment means x $t_{\alpha/2}$ for error degrees of freedom.

We have

$$\text{Var}\left(\bar{y}_i - \bar{y}_{.j}\right) = \frac{\sigma_e^2}{n_i} + \frac{\sigma_e^2}{n_j} = \sigma_e^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

$$\text{Standard Error } \left(\bar{y}_i - \bar{y}_{.j}\right) = \sigma_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

Hence, the critical difference (CD) for $\left(\bar{y}_i - \bar{y}_{.j}\right)$

$$= t_{\alpha/2}\left(\text{for error df}\right) x \left[\text{MSSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{1/2}$$

Since MSSE provides an unbiased estimate of σ_E^2.

If each treatment is replicated n times, that is nᵢ = n for i=1, 2, ..., k then

$$\text{CD for difference of mean } = t_{\alpha/2}\left(\text{for error df}\right) x \left[\text{MSSE}\left(\frac{2}{n}\right)\right]^{1/2}$$

12.5.3. Least Square Estimates of Effects:

The completely randomised model in equation (1) in Sub-section 9.5.2 is a fixed effect model. Proceeding exactly as in Section 6.4 of Unit 6, we shall get

$$\hat{\mu} = \frac{y_{..}}{n} = \bar{y}_{..} \text{ and } \hat{\alpha}_i = \bar{\alpha}_i = \bar{y}_{i.} - \bar{y}_{...} \qquad (2)$$

12.5.4. Variance of the Estimates:

Proceeding exactly as in Section 6.7 of Unit 6, we shall get

$$\text{Var}\left(\hat{\mu}\right) = \frac{\sigma_e^2}{n}; \text{ where } n = \sum_{i=1}^{k} n_i \qquad (3)$$

$$\text{And Var}\left(\alpha_i\right) = \text{Var}\left(\alpha_i\right) = \sigma_e^2 \left(\frac{1}{n_i} - \frac{1}{\sum_{i=1}^{k} n_i} \right) \qquad (4)$$

If we assume that each treatment is replicated an equal number of times i.e., if

$n_i = n$, (say), $i = 1, 2, \ldots, k$; then $n = \sum_{i=1}^{k} n_i = nk$

Hence, from equations (3) and (4), we get

$$\text{Var}\left(\hat{\mu}\right) = \frac{\sigma_e^2}{nk} \text{ and } \text{Var}\left(\hat{\alpha}_i\right) = \text{Var}\left(\sigma_i\right) = \sigma_e^2 \left(\frac{k-1}{nk} \right) \qquad (5)$$

12.5.5. Expectation of Sum of Squares:

Proceeding exactly as in Section 6.7 of Unit 6 [fixed effect model for one-way classified data], we get

$$E\left(SST\right) = E\left[\sum_{i=1}^{k} n_1 \left(\bar{y}_i - \bar{y}_{..}\right)^2 \right] = \left(k-1\right)\sigma_e^2 + \sum_{i=1}^{k} n_i \alpha_1^2$$

$$E\left(MSST\right) = E\left[\frac{S_T^2}{\left(k-1\right)} \right] = \sigma_e^2 + \frac{1}{k-1} \sum_{i=1}^{k} n_i \alpha_1^2$$

$$E\left(SSE\right) = \left(n-k\right)\sigma_e^2$$

$$E\left(MSSE\right) = E\left[\frac{S_E^2}{\left(n-k\right)} \right] = \sigma_e^2$$

The method of analysis of completely randomized design would be similar to one-way ANOVA, which has been illustrated below with the following example:

Example 1:

A person wanted to purchase a lot of electric drills. He got quotations from five manufacturers. For the selection, he wanted to conduct an experiment to estimate the time taken by each making a hole in a metallic sheet. As the sheet might not be uniform all over in respect of thickness and hardness, he marked 20 places on the sheet and applied four drills from each concern in 4 randomly selected places to make holes. The time for making each hole was recorded and these formed the observations. The observations in seconds are shown below in brackets along with marks of the drills denoted by D_1, D_2, D_3, D_4 and D_5.

$D_1(19)$ $D_3(22)$ $D_4(20)$ $D_1(20)$

$D_5(29)$ $D_2(24)$ $D_5(30)$ $D_3(24)$

$D_2(26)$ $D_4(25)$ $D_1(16)$ $D_2(22)$

$D_5(28)$ $D_3(25)$ $D_5(31)$ $D_4(28)$

$D_4(27)$ $D_1(16)$ $D_2(27)$ $D_3(20)$

Conduct the experiment by adopting a completely randomized design.

Solution: The analysis of the given design is done by one-way analysis of variance method. The data is analyzed and computation results are given as below:

The totals of time records for 4 holes by each of the different makes are denoted by T_1, T_2, T_3, T_4 and T_5 are shown below.

$$T_1 = 71, T_2 = 99, T_3 = 91, T_4 = 100, T_5 = 118$$

Grand Total (G) = 479

Correction Factor (CF) $= \frac{G^2}{N} = \frac{(479)^2}{20} - 11472.05$

Total Sum of Squares (TSS) $= 21^2 + 18^2 + 22^2 + \ldots + 31^2 + 20^2 - 11472.05$

$$= 11847 - 11472.05 = 374.95$$

Sum of Squares due to Makes (SSM)

$$= \frac{(71)^2 + (99)^2 + (91)^2 + (100)^2 + (118)^2}{4} - 11472.05$$

$$= 11761.75 - 11472.05 = 289.70$$

Sum of Squares due to Error (SSE) = TSS – SSM

$$= 374.95 - 289.70 = 85.25$$

Analysis of Variance Table

Sources of Variation	**DF**	**SS**	**MSS**	**F**
Makes	4	289.7	72.425	12.75
Error	15	85.25	5.68	
Total	19	374.95		

The tabulated value of F at 1 per cent level of significance for 4 and 15 df is 4.89. Thus, the calculated value of F viz. 12.75 shows that Make to Make variation is highly significant thereby indicating that the hypothesis that the time periods taken by the different Makes in boring a hole are, on an average, the same, is rejected. So multiple comparison test will be applied for different Makes.

Mean for Different Makes:

Makes				
$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	$\bar{D}_4$	$\bar{D}_5$
17.74	24.75	22.75	25.00	29.50

$$SE = SE = \sqrt{\frac{2MSSE}{n}} = \sqrt{\frac{2x5.68}{4}} = 1.69$$

Critical difference at 1% level of significance

CD = $t_{\alpha/2}$ (for error df) x SE = 3.055 x 1.69 = 5.16

The initial difference indicates that the Make D5 is significantly better than all the other Makes.

Pair of Treatments	**Difference**	**CD**	**Inference**
D_1, D_2	$\left\|\overline{D_1}-\overline{D_2}\right\|$=7.10	5.16	Significant
D_1, D_3	$\left\|\overline{D_1}-\overline{D_3}\right\|$=5.10	5.16	Insignificant
D_1, D_4	$\left\|\overline{D_1}-\overline{D_4}\right\|$=7.26	5.16	Significant
D_1, D_5	$\left\|\overline{D_1}-\overline{D_5}\right\|$=11.26	5.16	Significant
D_2, D_3	$\left\|\overline{D_2}-\overline{D_3}\right\|$=2.00	5.16	Insignificant
D_2, D_4	$\left\|\overline{D_2}-\overline{D_4}\right\|$=0.25	5.16	Insignificant
D_2, D_5	$\left\|\overline{D_2}-\overline{D_5}\right\|$=4.75	5.16	Insignificant
D_3, D_4	$\left\|\overline{D_3}-\overline{D_4}\right\|$=2.25	5.16	Insignificant
D_3, D_5	$\left\|\overline{D_3}-\overline{D_5}\right\|$=6.75	5.16	Significant
D_4, D_5	$\left\|\overline{D_4}-\overline{D_5}\right\|$=4.5	5.16	Insignificant

12.6. SUITABILITY OF CRD:

The following are some situations, in which one can apply the complete randomised design:

1) The CRD is used in the situations where experimental materials are homogeneous. That is why, CRD is mostly used in chemical, biological and banking experiments, where the experimental material is thoroughly mixed powder, liquid or chemical.

2) The CRD is used in the situations where the observations on some units are missing or destroyed. This feature of missing observation does not disturb the analysis of the design.

3) In agricultural experiments, this design is not used because experimental material is not homogeneous.

12.6.1. Advantages and Disadvantages of CRD:

Advantages of CRD

1) In this design any number of treatments and replications can be used. There may be different number of replications for different treatments.

2) Analysis is simple and easy even if the number of replications is unequal for each treatment. In such case experimental error will differ from treatment to treatment.

3) If some of the observations are missing or destroyed or not available due to some reasons, the analysis can be done without any problem.

4) It provides large degree of freedom for error sum of squares. This increases the sensitivity of the experiment.

5) In CRD there is no condition on the number of replications of the treatments, they can be increased or decreased according to the need of the experimenter. Thus, the design is flexible.

Disadvantages of CRD:

1) The main disadvantage of CRD is that the principle of local control has not been used in this design. Due to this fact, the experimental error is inflated. This is the main reason for the criticism of CRD.

2) In agricultural experiments, the design is seldom used because the experimental material is not homogenous.

12.7. SUMMARY:

In this unit, we have discussed:

1) The experimental design;

2) The planning and classification of experimental designs;

3) Principles of design of experiments;

4) Completely randomised design;

5) Layout of CRD;

6) The statistical analysis of CRD; and

7) Advantages and disadvantages as well as suitability of CRD.

Example 2: Carryout the ANOVA for the given following data of yields of 5 varieties, 7 observations on each variety:

Variety	**Observations**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**13**	**15**	**14**	**14**	**17**	**15**	**16**
2	**11**	**11**	**10**	**10**	**15**	**9**	**12**
3	**10**	**13**	**12**	**15**	**14**	**13**	**13**
4	**16**	**18**	**13**	**17**	**19**	**14**	**15**
5	**12**	**12**	**11**	**10**	**12**	**10**	**10**

The analysis of the given design is done by one-way analysis of variance method. The data is analyzed and computation results are given as below:

Correction Factor (CF) = 6072.03

Raw Sum of Squares (RSS) = 6293

Total Sum of Squares (TSS) = 220.97

Sum of Squares due to Variety (SSV) = 138.40

Sum of Squares due to Error (SSE) = TSS – SSV

= 220.97-138.40 = 82.57

ANOVA TABLE

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ratio**	
				Calculated	**Tabulated**
Variety	4	138.40	34.60	12.58	2.66
Error	30	82.57	2.75		
Total	34	220.97			

Null Hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_5$

Since, calculated value of F is greater than the tabulated value of F, we reject the null hypothesis and conclude that variety effects are significantly different.

Mean for Different Varieties:

Varieties				
$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	$\bar{D}_4$	$\bar{D}_5$
14.86	11.14	12.86	16.00	11.00

$$SE = SE = \sqrt{\frac{2MSSE}{n}} = \sqrt{\frac{2x13.34}{7}} = 1.95$$

Critical difference at 1 % level of significance

= $t_{\alpha/2}$ (for error df) x SE = 3.055 x 1.95 = 5.96

The initial difference indicates that the Variety D4 is significantly better than all the other Varieties.

Pair of Treatments	Difference	CD	Inference
D_1, D_2	$\lvert \overline{D_1} - \overline{D_2} \rvert$ = 3.72	5.96	Insignificant
D_1, D_3	$\lvert \overline{D_1} - \overline{D_3} \rvert$=2.00	5.96	Insignificant
D_1, D_4	$\lvert \overline{D_1} - \overline{D_4} \rvert$=1.14	5.96	Insignificant
D_1, D_5	$\lvert \overline{D_1} - \overline{D_5} \rvert$=3.86	5.96	Insignificant
D_2, D_3	$\lvert \overline{D_2} - \overline{D_3} \rvert$=1.72	5.96	Insignificant
D_2, D_4	$\lvert \overline{D_2} - \overline{D_4} \rvert$=4.86	5.96	Insignificant
D_2, D_5	$\lvert \overline{D_2} - \overline{D_5} \rvert$=0.14	5.96	Insignificant
D_3, D_4	$\lvert \overline{D_3} - \overline{D_4} \rvert$=3.14	5.96	Insignificant
D_3, D_5	$\lvert \overline{D_3} - \overline{D_5} \rvert$=1.86	5.96	Insignificant
D_4, D_5	$\lvert \overline{D_4} - \overline{D_5} \rvert$=5.00	5.96	Insignificant

12.8. SELF-ASSESSMENT QUESTIONS:

1) Explain Principles of Design of Experiments
2) Explain Layout of Completely Randomized Design
3) Explain the Statistical Analysis of Completely Randomized Design
4) Explain the Least Square Estimates of Effects
5) Explain the Suitability of CRD

12.9 SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.
3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

Dr. B. Hari Mallikarjuna Reddy

LESSON-13

RANDOMISED BLOCK DESIGN

13.0. OBJECTIVES:

After studying this unit, you would be able to

- Explain the randomized block design;
- Describe the layout of RBD;
- Explain the statistical analysis of RBD;
- Find out the missing plots in RBD; and
- Explain the advantages and disadvantages as well as the suitability of RBD.

STRUCTURE:

13.1. INTRODUCTION:

The completely randomized design was simple due to the reason that principle of local control was not used and it was assumed that the experimental material is homogeneous, but it is observed that the experimental material is not fully homogeneous. In agricultural field experiments sometimes, a fertility gradient is present in one direction. In such situation the simple method of controlling variability of the experimental material consists in stratifying or grouping the whole experimental area into relatively homogeneous strata or sub-groups (called blocks), perpendicular to the direction of fertility gradient. These blocks are so formed that plots within a block are homogeneous and between blocks are heterogeneous. In other words, there may be less variation within a block and major difference or variation between blocks. It is to be kept in mind that familiarity with the nature of experimental units is necessary for an effective blocking of the material. The procedure of division of experimental material into a number of blocks give rise to a design known as Randomized block design (RBD) which can be defined as an arrangement of t treatments in r blocks such that each treatment occurs precisely once in each block.

In other words, when the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that those experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units. Within each group (or block). This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks is smaller because of homogeneous blocks.

13.2. LAYOUT OF RANDOMISED BLOCK DESIGN:

The entire experimental material is divided into a number of blocks equal to the number of replications for each treatment. Then each block is divided into a number of plots equal to the number of treatments. For example, if we have 4 treatments A, B, C and D and each treatment is to be replicated 3 times. Then according to the condition of RBD, we will arrange the experimental material in three blocks each of size 4, i.e. each block consists of 4 plots. After arranging the experimental material into a number of blocks, treatments are allocated to each block separately. That is randomization is applied afresh for each block and thus, it will be independent for each block. The method is illustrated below by the following arrangement of 3 blocks and 4 treatments:

Layout of RBD with 4 Treatments

Block I	A	B	D	C
Block II	C	A	D	B
Block III	D	B	C	A

13.3. STATISTICAL ANALYSIS OF RBD:

If in RBD a single observation is made on each of the experimental units, then its analysis is analogous to ANOVA for fixed effect model for a two-way classified data with one observation per cell and the linear model effects to be (additive) becomes

$$y_{ji} = \mu + \alpha_i + \beta_j + e_{ij}; \qquad i = 1, 2, ..., P; j = 1, 2, ..., q.$$

where, y ij is the yield or response of the experimental unit receiving the ith treatment in the jth block, μ is the general mean effect, α_i is the effect due to the ith treatment, β_j in the effect due to jth block or replicate and mathfrak e_{ij} is identically and independently distributed i.e. e_{ij} follows (i.i.d.) N (0, σ_e^2),

where $\mu\ \alpha_i$ and β_j are constants so that $\sum_{i=1}^{p}\alpha_i = 0$ and $\sum_{j=i}^{q}\beta j = 0.$

If we write that

$$\sum_i \sum_j yij = y.. = G = \text{grand total of all the pxq observations.}$$

$$\sum_j yij = y_i = \alpha_i = \text{ Total for i}^{th} \text{ treatmetn}$$

$$\sum_j yij = y_i = \beta_i = \text{ Total for j}^{th} \text{ block}$$

Then heuristically, we get

$$= q\sum_i (\bar{y}_i - \bar{y}..)^2 + p\sum_j (\bar{y}_j - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}._j + \bar{y}..)^2$$

$$TTS = \sum_i \sum_j (y_{ij} - \bar{y}..)^2$$

$$SST = q\sum_i (\bar{y}_i - \bar{y}_{..})^2 = S_T^{\ 2}(\text{say})$$

$$SSB = p\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = S_B^{\ 2}$$

$$SSE = S_E^{\ 2} - TSS - SSB = SST$$

$$F_B = \frac{MSST}{MSSE}$$

The product terms vanish since the algebraic sum of deviations form mean is zero. This

TSS =SSE+ SSB+ SST

Where TSS, SST, SSB and SSE are that total sum of squares, sum of squares due to treatments (between treatments SS), of squares due to blocks and sum of squares due to error (i.e., within treatment SS) given respectively by

$$TTS = \sum_{i}\sum_{j}(y_{ij} - \bar{y}..)^2$$

$$SST = q\sum_{i}(\bar{y}_{i} - \bar{y}_{..})^2 = S_T^{\,2}(\text{say})$$

$$SSB = p\sum_{j}(\bar{y}_{.j} - \bar{y}_{..})^2 = S_B^{\,2}$$

$$SSE = S_E^{\,2} - TSS - SSB = SST$$

Hence, the total sum of squares is partitioned three sums of squares whose degree of freedom make the total to the degree freedom of TSS.

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ration (F)**
Treatments	P^{-1}	S_T^2	MSST=S_T^2/(P-1)	$F_T = \dfrac{MSST}{MSSE}$
Blocks	q-1	S_B^2	MSST=S_B^2/(q-1)	$F_B = \dfrac{MSST}{MSSE}$
Error	(p-1) (q-1)	S_E^2	MSST=S_E^2/ (P-1)(q-1)	
Total	Pq-1			

Under the null hypothesis, H$_0$: $\alpha_1 = \alpha_2 = ... = \alpha_p$ against the alternative that all α's are not equal, the test statistic

$$F_T = \frac{MSST}{MSSE} \qquad \text{Follows F[(P-1) (q-1)]}$$

i.e., F_T follows F-distribution with [(P-1), (P-1) (q-1) df.

i.e., $F_T \leq F$ with [(p-1), (p-1) (q-1)] df at α level of significance, (Usually 5%) then H$_0$ is rejected and we conclude that treatments differ significantly.

If $F_T <$ F with [(p-1), (p-1)] df at α level of Significance, then H_0 may be accepted, i.e. the data do not provide any evidence against the null hypothesis which may be accepted.

Similarly, under the null hypothesis, $H_0 : \beta_1 = \beta_2 = ... = \beta_q$ against the alternative that all β 's are not equal, the test statistic.

$$F_T = \frac{\text{MSSB}}{\text{MSSE}} \qquad \text{follows } F[(q-1), (p-1)\ (q-1)]$$

13.3.1. Least Square Estimates of Effects:

Proceeding exactly similar as in CRD, and replacing K by p, n by q and taking N=pq, the estimates of the parameters μ, α_i and β_j are given by:

$$\bar{\mu} = \bar{y}..., \hat{a}_i = \bar{y}_{i.} - \bar{y}_{..}, \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..} \qquad …….. (1)$$

13.3.2. Variance of the Estimates:

Proceeding exactly similar as in CRD, we shall get

$$\text{Var}\ (\hat{\mu}) = \frac{\sigma_e^2}{pq}$$

$$\text{Var}\ (\hat{\alpha}_i) = \frac{(p-1)}{pq} \sigma_e^2$$

and $$\text{Var}\ (\hat{\beta}_j) = \frac{(q-1)}{pq} \sigma_e^2$$

13.3.3. Expectation of Sum of Squares:

Proceeding exactly as in CRD, we get

$$E[SST] = (p-1)\sigma_e^2 + q\sum_i \alpha_i^2$$

$$E\left[\frac{(SST)}{(p-1)}\right] = E(MSST) = \sigma_e^2 + \frac{q}{(p-1)}\sum_i \alpha_i^2$$

$$E(SSB) = (q-1)\sigma_e^2 + P\sum_j \beta_j^2$$

$$E\left[\frac{(SST)}{(q-1)}\right] = E(MSSB) = \sigma_e^2 + \frac{q}{(q-1)}\sum_j \beta_j^2$$

$$E(SSE) = (q-1)(p-1)\sigma_e^2$$

$$E\left[\frac{(SSE)}{(q-1)(p-1)}\right] = E(MSSE) = \sigma_e^2$$

Hence under the null hypothesis

$$H_0\beta : \alpha_1 = \alpha_2 .... = \alpha_p = 0;$$

$$H_0\beta : \beta_1 = \beta_2 .... = \beta_p = 0;$$

$$E(MSST) = \sigma_e^2 \text{ and } E(MSSB) = \sigma_e^2$$

i.e. each of the mean sum of squares due to treatments and blocks gives an unbiased estimate of the error variance σ_e^2 under the null hypothesis $H_{0\alpha}$ and $H_{0\beta}$ respectively.

Example 1:

There were 4 different makes of cars. A problem was posed to estimate the petrol consumption rates of the different makes of cars for suitable average speed and compare them. The following experiment could be conducted for an inference about the problem:

Five different cars of each four makes were chosen at random. The five cars of each make were put on road on 5 different days. The cars of A make run with different speeds on different days. The speeds were 25, 35, 50, 60 and 70 mph. Which car was to put on the road on which day and what speed it should have was determined through a chance mechanism subject to the above conditions of the experiment. The procedure was adopted for each of the makes of cars. For each car, the number of miles covered per gallon of petrol was observed. The observations are presented below:

TABLE: MILES PER GALLON OF PETROL

Makes of Car	**Speed of the Cars in Miles Per Hour (mph)**						**Average**
	25	**35**	**50**	**60**	**70**	**Total**	
A	20.6	19.5	18.1	17.9	16.0	92.1	18.42
B	19.5	19.0	15.6	16.7	14.1	84.9	16.42
C	20.5	18.5	16.3	15.2	13.7	84.2	16.84
D	16.2	16.5	15.7	14.8	12.7	75.9	15.18
Total	76.8	73.5	65.7	64.6	65.5	337.1	

Carry out the analysis of the given RBD.

Solution:

Here the makes of the cars are the treatments and the other controlled factor is the speed, the variance for which has been eliminated through the design which is thus actually a randomized block design with the speeds as blocks. The specific cars used, the effects of the days, drivers and possibly some other effects contributed to the error variance.

Here,

Correction Factor (CF) = $\frac{(337.1)^2}{20} = 5681.82$

Raw Sum of Squares = $(20.6)^2+(19.5)^2+\ldots+(13.7)^2 + (12.7)^2$ =5781.41

Total Sum of Squares (TTS) = 5681.41.5881.82=99.59

Sum of Squares due to Speed (SSS)

$$= \frac{(76.8)^2 + (73.5)^2 + \ldots.. + (64.6)^2 + (56.5)^2}{4} - \text{CF}$$

=66.04

Sum of Squares due to Makes (SSM)

$$= \frac{(92.1)^2 + (84.9)^2 + (84.2)^2 + (75.9)^2}{5} - \text{CF}$$

=28.78

Sum of Squares due to Errors (SSE)

= TSS-SSS-SSM

=99.59-66.04-28.78=4.77

ANALYSIS OF VARIANCE TABLE

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ration**	
				Calculated	**Tabulated**
Speeds	4	66.04	16.57	41.27	3.26
Treatments (Makes)	3	28.78	9.59	23.97	3.49
Error	12	4.77	0.40		
Total	**19**	**99.59**			

In both the cases either for speeds or for makes, calculated value of F is greater than tabulated value of F at 5% level of significance and thus null hypothesis is rejected.

In the above experiment, we are interested only on makes so multiple comparison test will be applied for different makes.

Mean number of miles per gallon for different Makes

Makes

$\bar{A}$	$\bar{B}$	$\bar{C}$	$\bar{D}$
18.42	**16.98**	**16.84**	**15.18**

$$SE = \sqrt{\frac{2MSSE}{5}} = \sqrt{\frac{2x0.40}{5}} = 0.40$$

Critical difference at 1 % level of significance

CD $t_{\alpha/2}$ (for error df) x SE=3.055x 0.40 =122

The initial difference indicates that the Make A is significantly better than all the other Makes.

Pair of Treatments	Difference	CD	Inference
A,B	$\left\|\bar{A}-\bar{B}\right\|=1.44$	1.22	Significant
A,D	$\left\|\bar{A}-\bar{C}\right\|=1.58$	1.22	Significant
A,C	$\left\|\bar{A}-\bar{D}\right\|=3.24$	1.22	Significant
B,C	$\left\|\bar{B}-\bar{C}\right\|=0.14$	1.22	Significant
B,D	$\left\|\bar{B}-\bar{D}\right\|=1.8$	1.22	Significant
C,D	$\left\|\bar{C}-\bar{D}\right\|=1.66$	1.22	Significant

Example 2:

Carryout the analysis of the following design:

Varieties	Blocks			
	I	II	III	IV
A	7	16	10	11
B	14	15	15	14
C	8	16	7	11

Varieties	Blocks				Total
	I	II	III	IV	
A	7	16	10	11	44
B	14	15	15	14	58
C	8	16	7	11	102
Total	**29**	**74**	**32**	**36**	**144**

Correction Factor (CF) $= \frac{(144)^2}{12} = 1728$

Raw Sum of Squares (RSS) $= (7)^2 + (14)^2 + ... + (14)^2 + (11)^2 = 1858$

Total Sum of Squares (TSS) =1858-1728=130

Block Sum of Squares (SSB) $= \frac{(29)^2 + (47)^2 + (32)^2 + (36)^2}{3} - \text{CF}$

$$= \frac{841 + 2209 + 1024 + 1296}{3} - 1728$$

$$= 1790 - 1728 = 62$$

Variety Sum of Squares (SSV) $= \frac{(44)^2 + (58)^2 + (42)^2}{4} - \text{CF}$

$$= \frac{1936 + 3364 + 1764}{4} - 1728$$

$$= 1766 - 1728 = 38$$

Sum of Squares due to Error (SSE) $= \text{TSS} - \text{SSV} - \text{SSB}$

$$=130 - 62 - 39 = 30$$

ANOVA TABLE

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ration**	
				Calculated	**Tabulated**
Variety	2	38	19	3.8	5.14
Blocks	3	62	20.67	4.13	4.76
Error	6	30	5		
Total	**11**	**130**			

In both these cases either for varieties or for blocks, calculated value of F is less than tabulated value off at 5% level of significance and thus null hypothesis is accepted and inferred that variety effect and block effect are insignificant.

13.4. MISSING PLOTS TECHNIQUE IN RBD:

Sometimes observations from one or more experimental units are not found (missing) due to some unavoidable causes. There may be some unforeseen causes for example in agricultural experiments damage by animal or pets, in animal experiment any animal may die or observations from one or more plot is excessively large as compared to other plots and thus accuracy of such observation is often in doubt. In such observation is often in doubt. In such situations, these observations are omitted and treated as missing.

In case of missing observations, analysis is done by estimating the missing observation. This type of analysis was given by Yates (1937) and it is known as missing plot technique.

13.4.1. One Missing Plot:

Suppose without loss of generality that observation for treatment 1 in block 1 i.e. y_{11} is missing and let it is Y, then the observations for a RBD may be represented as below:

	T_1	T_2	…..	T_i	….	T_p	Total
B_1	$Y_{11}=Y$	Y_{21}	….	Y_{i1}	….	Y_{p1}	$B_1'+Y$
B_2	Y_{12}	Y_{22}	….	Y_{i2}	….	Y_{p2}	B_2
…	…	…	…	…	…	…	…
B_j	Y_{1j}	Y_{2j}	…	Y_{ij}	…	Y_{pj}	B_j
…	…	…	…	…	…	…	…
B_q	Y_{1q}	Y_{2q}	…	Y_{iq}	…	Y_{pq}	B_q
Total	$T_1'+Y$	T_2	…	T_i	…	T_p	$G'+Y$

Where,

B_1'=total of all available (p-1) observations in 1st block

T_1'=total of all available (q-1) observations in 1st treatment.

G'=total of all available (pq-1) observations

On the basis of these totals we calculate different SS's as follows:

Sum of Squares for blocks (SSB) = $\dfrac{\left(B_1^{'}+Y\right)^2+\sum\limits_{j=2}^{q}B_j^2}{p}-\dfrac{\left(G'+Y\right)^2}{pq}$

Sum of Squares for Treatments (SST) = $\dfrac{\left(T_1^{'}+Y\right)^2+\sum\limits_{j=2}^{q}T_j^2}{q}-\dfrac{\left(G'+Y\right)^2}{pq}$

Total Sum of Squares (TSS) = $\sum\limits_{i}\sum\limits_{j}y_{ij}^2+Y^2-\dfrac{\left(G'+Y\right)^2}{pq}$ where $(i,j)\neq(1,1)$

Sum of Squares due to Error (SSE) = TSS – SSB – SST

$$SSE = Y^2+\frac{\left(G'+Y\right)^2}{pq}-\frac{\left(B_1^{'}+Y\right)^2}{p}-\frac{\left(T_1^{'}+Y\right)^2}{q} + \text{terms not involving Y}$$

For obtaining the value of Y, we minimize the sum of squares due to error with respect to Y. this is obtained by solving the equation.

$$\frac{\partial(SSE)}{\partial Y}=2Y+\frac{2(G'+y)}{qp}-\frac{2\left(B_1^{'}+Y\right)}{p}-\frac{2\left(T_1^{'}+Y\right)}{q}=0$$

$$=Y+\frac{Y}{pq}-\frac{Y}{p}-\frac{Y}{q}=\frac{T_1^{'}}{q}+\frac{B_1^{'}}{p}-\frac{G'}{pq}$$

$$=\frac{Y(pq+1-q-p)}{pq}=\frac{pT_1^{'}+qB_1^{'}-G'}{pq}$$

$$\hat{Y}=\frac{pT_1^{'}+qB_1^{'}-G'}{(p-1)(q-1)}$$

$\hat{Y}$ is the least square estimate of the yield of the missing plot. The value of Y is inserted in the original table of yield and ANOVA is performed in the usual way except that for each missing observation 1 df is subtracted from total and consequently from error df.

13.4.2. Two Missing Plots:

For two missing values, we convert the problem into one missing value by putting any value say the overall mean or mean of the available values of that block for which one value is missing or mean of the available values of that replicate in any missing cell and obtain the estimate of the second missing value by the above prescribed estimation formula. Then we put the estimate of this second missing value and estimate the first missing value for which originally mean was taken. We go on repeating the same procedure until we obtain two successive estimates which are not materially different. Method is illustrated below with examples.

Example 3:

In the following data two values are missing. Estimate these values by Yates method and analyse:

Treatments	**Blocks**		
	I	**II**	**III**
A	12	14	12
B	10	y	8
C	x	15	10

Solution: We convert the two missing plots problems into one missing plot problem, for which we take the average of the values of I block in which x is missing. This average is (10+12)/2=11. Thus, the estimate of x is taken to be $x_1=11$ and it is inserted in place of x and form the following table of totals:

Treatments	**Blocks**			**Total**
	I	**II**	**III**	
A	12	14	12	$T_A=38$
B	10	Y	8	$T_B=18+Y$
C	11	15	10	$T_C=36$
Total	$B_1=33$	$B_2=29+y$	$B_3=30$	$G=92+y$

Thus, from the above table we get

P=3, q=3, B_2'=29, T_B'=18, G'=92

Applying the missing estimation formula

$$\hat{y} = \frac{pT_1^{'} + qB_1^{'} - G}{(q-1)(p-1)} = \frac{3\times18+3\times29-92}{4}$$

$$= \frac{54+87-92}{4} = \frac{49}{4} = 12.25 \approx 12$$

Now the estimated value of y is taken to be y_1 = 12 and it is inserted in place of y and the following table of totals is formed by taking x unknown:

Treatments	**Blocks**			**Total**
	I	**II**	**III**	
A	12	14	12	T_A=38
B	10	12	8	T_B=30
C	x	15	10	T_C=25+x
Total	B_1=22+x	B_2=41	B_3=30	G=93+x

Thus, from the above table we get p = 3, q = 3, B_1'=22, T_C' = 25, G' = 93

Again, applying the missing estimation formula

$$\hat{x} = \frac{3\times25+3\times22-93}{4}$$

$$= \frac{75+66-93}{4} = \frac{48}{4} = 12$$

Thus, x_2=12

Again, using x_2=12, we estimate the second estimate of y i.e. y_2 for which

$B_2^{'} = 29,\ T_B^{'} = 18,\ G' = 92$

$$\hat{y} = \frac{3\times 18 + 3\times 29 - 93}{4}$$

$$= \frac{54+87-93}{4} = \frac{47}{4} = 11.75 = 12$$

We see that the second estimate of y i.e. y_2 is not materially different from y_1.

Thus, we take the estimated values of $\hat{x} = 12$ and $\hat{y} = 12$. Inserting both the estimated values of x and y we get the following observations:

Treatments	**Blocks**			**Total**
	I	**II**	**III**	
A	12	14	12	T_A=38
B	10	12	8	T_B=30
C	12	15	10	T_C=37
Total	B_1=34	B_2=41	B_3=30	G=105

Correction Factor (CF) = $\frac{(105)^2}{9} = \frac{11025}{9} = 1225$

Raw Sum of Square (RSS) = $(12)^2 + (10)^2 + ... + (8)^2 + (10)^2 = 1261$

Total Sum of Squares (TSS) = 1261-1225=36

Treatment Sum of Squares (SST) = $\frac{(38)^2 + (30)^2 + (37)^2}{3} - CF$

$$= \frac{1444 + 900 + 1369}{3} - 1225$$

$$= \frac{3713}{3} - 1225 = 1237.67 - 1225$$

$$= 12.67$$

Block Sum of Squares (SSB) = $\frac{(34)^2 + (41)^2 + (30)^2}{3} - \text{CF}$

$$= \frac{1156 + 1681 + 900}{3} - 1225$$

$$= 1245.67 – 1225 = 20.67$$

Error Sum of Squares (SSE) = TSS – SST – SSB

$$= 36 – 12.67 – 20.67 = 2.66$$

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ratio**	
				Calculated	**Tabulated**
Treatments	3-1=2	12.67	6.34	4.77	9.55
Blocks	3-1=2	20.67	10.34	7.77	9.55
Error	4-2=2	2.66	1.33		
Total	**8-2=6**				

In case of both treatments and blocks, calculated value of F is less than tabulated value of F at 5% level of significance, thus treatment and block means are not significantly different.

13.5 SUITABILITY OF RBD:

1) The RBD is suitable in the situations where it is possible to divide the experimental material into a number of blocks. If it is not possible to divide the experimental material, RBD cannot be used.

2) the RBD is suitable only when the number of treatments is small because as the number of treatments increases, the block size also increases and it disturbs the homogeneity of the block.

3) RBD is suitable only when experimental material is heterogeneous with respect to one factor only. If there is two-way heterogeneity, LSD is used.

13.5.1. Advantages and Disadvantages of RBD:

Advantages of RBD:

The RBD has many advantages over other designs. Some of them are listed below:

1) It is a flexible design. It is applicable to moderate number of treatments. If extra replication is necessary for some treatment, this may be applied to more than one unit (but to the same number of units) per block.

2) Since all the three principles of design of experiments are used, the conclusions drawn from RBD are more valid and reliable.

3) If data from individual units be missing then, analysis can be done by estimating it.

4) This is the most popular design in view of its simplicity, flexibility and validity. No other design has been used so frequently as the RBD.

5) This design has been shown to be more efficient or accurate than CRD, for most types of experimental work. The elimination of block sum of squares from error sum of squares, usually results in a decrease of error sum of squares.

6) Analysis is simple and rapid.

Disadvantages of RBD:

1) The main disadvantage of RBD is that if the blocks are not internally homogeneous, then a large error term will result. In field experiments, it is usually observed that as the number of treatments increases, the block size increases and so one has lesser control over error.

2) The number of replications for each treatment is same. If replication is not same, the only remedy is to adopt CRD.

3) It cannot control two-sided variation of experimental material simultaneously. That is why, it is not recommended when experimental material contains considerable variability.

13.6. SUMMARY:

In this unit, we have discussed:

1) The randomised block design;

2) The layout of RBD;
3) The statistical analysis of RBD;
4) The missing plot techniques in RBD; and
5) The advantages and disadvantages as well as the suitability of RBD.

Example 4: Carryout the analysis of following design:

Block			
I	**II**	**III**	**IV**
A **8**	**C** **10**	**A** **6**	**B** **10**
C **12**	**B** **8**	**B** **9**	**A** **8**
B **10**	**A** **8**	**C** **10**	**C** **9**

The given design is solved by method of analysis of variance for two-way classified data. The computation results are given as follows:

Correction Factor (CF) = 972

Raw Sum of Squares (RSS) = 998

Total Sum of Squares (TSS) = 26

Block Sum of Squares (SSB) = 4.67

Treatment Sum of Squares (SST) = 15.5

Error Sum of Squares (SSE) = 5.83

ANOVA TABLE

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Variety	2	15.5	7.7	7.94	5.14
Blocks	3	4.67	1.56	1.61	4.76
Error	6	5.83	0.97		
Total	**11**	**26**	**32**		

In case of variety, calculated value of F is greater than the tabulated value at F at 5% level of significance, so we reject the null hypothesis and conclude that the treatment effect is significant, while for blocks, it is not significant. For pairwise testing, we have to find the standard error of difference of two treatment means:

$$SE\sqrt{\frac{2MSSE}{q}}=\sqrt{\frac{2x0.97}{3}}=0.80$$

Critical Difference (CD) = SE x $t_{\alpha/2}$ at error df

= 0.80x2.447=1.96

Treatment means are

$$\bar{A}=\frac{30}{4}=7.5, \qquad \bar{B}=\frac{37}{4}=9.25 \qquad \bar{C}=\frac{41}{4}=10.25$$

Pair of Treatments	Difference	CD	Inference
A,B	$\lvert\bar{A}-\bar{B}\rvert=1.76$	1.96	Insignificant
A,C	$\lvert\bar{A}-\bar{C}\rvert=2.75$	1.96	Significant
B,C	$\lvert\bar{B}-\bar{C}\rvert=1.00$	1.96	Insignificant

Example 5: For the given data the yield of the treatment C in 2nd block is missing.

Estimate the missing value and analyse the data:

Blocks	**Treatments**			
	A	**B**	**C**	**D**
I	**105**	**114**	**108**	**109**
II	**112**	**113**	**Y**	**112**
III	**106**	**114**	**105**	**109**

We have p=3, q=4, B_3 = 213, T_2 =337, G=1207 and the value of

$\hat{y}$=109

Therefore,

Correction Factor = 144321.33

Raw Sum of Squares = 144442.00

Total Sum of Squares = 120.64

Treatment Sum of Squares = 76.67

Block Sum of Squares =20.67

Error Sum of Squares =23.33

ANOVA TABLE

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ratio**	
				Calculated	**Tabulated**
Treatments	3- 1=2	20.67	10.33	2.21	5.79
Blocks	4- = 3	76.67	25.55	5.48	5.41
Error	6- 1=3	23.33	4.66		
Total	11- 1=10	120.67	32		

In the above experiment, we are interested only treatments, so multiple comparison test will be applied for different treatments.

$$SE\sqrt{\frac{2MSSE}{q}} = \sqrt{\frac{2x4.66}{3}} = 1.76$$

$$CD = SExt_{\alpha/2} \text{ at error df}$$

$$= 1.76\,x2.447 = 4.31$$

Treatment means are

$$\bar{A} = \frac{323}{3} = 107.67, \quad \bar{B} = \frac{341}{3} = 113.67, \quad \bar{C} = \frac{322}{3} = 107.33, \quad \bar{D} = \frac{330}{3} = 110,$$

Pair of Treatments	Difference	CD	Inference
A,B	$\left\|\bar{A}-\bar{B}\right\| = 6.0$	4.31	Significant
A,D	$\left\|\bar{A}-\bar{C}\right\| = 0.3$	4.31	Insignificant
A,C	$\left\|\bar{A}-\bar{D}\right\| = 2.3$	4.31	Insignificant
B,C	$\left\|\bar{B}-\bar{C}\right\| = 6.3$	4.31	Significant
B,D	$\left\|\bar{B}-\bar{D}\right\| = 3.7$	4.31	Insignificant
C,D	$\left\|\bar{C}-\bar{D}\right\| = 2.7$	4.31	Insignificant

13.7. SELF-ASSESSMENT QUESTIONS:

1) Explain Layout of Completely Randomized Design
2) Explain the Statistical Analysis of Completely Randomized Block Design
3) Explain the Least Square Estimates of Effects of RBD

4) Missing Plots Technique in RBD
5) Explain the Suitability of RBD

13.8. SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.
2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.
3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

Dr. B. Hari Mallikarjuna Reddy

LESSON-14

LATIN SQUARE DESIGN

14.0. OBJECTIVES:

After studying this unit, you would be able to

- Explain the latin square design;
- Describe the layout of LSD;
- Explain the statistical analysis of LSD;
- Find out the missing plot in LSD; and
- Explain the advantages and disadvantages of LSD.

STRUCTURE:

14.1 INTRODUCTION

We know that RBD is used when experimental material is heterogeneous with respect to one factor and this factor of variation is eliminated by grouping the experimental material into a number of homogeneous groups called blocks. This grouping can be carried one step forward and we can group the units in two ways, each way corresponding to a source of variation among the units, and get the LSD. In agricultural experiments generally, fertility gradient is not always known and, in such situations, LSD is used with advantage. Then LS Deliminates the initial variability among the units in two orthogonal directions.

The Latin Square design represents, in some sense, the simplest form of a row-column design. It is used for comparing m treatments in m rows and m columns, where rows and columns represent the two blocking factors. Latin squares and their combinatorial properties have been attributed to Euler (1782). They were proposed as experimental designs by Fisher (1925, 1926), although De Palluel (l788) already utilized the idea of a 4x4 latin square design for an agricultural experiment (see Street and Street, 1987, 1988).

14.2. LAYOUT OF LATIN SQUARE DESIGN (LSD):

Mathematically speaking, the latin square of order m is an arrangement of m latin letters in a square of m rows and m columns such that every latin letter occurs once in each row and once in each column, or more generally, the arrangement of m symbols in a m x m array such that each symbol occurs exactly once in each row and column. In the context of experimental design, the latin letters are the treatments. Latin squares exist for every m. A reduced latin square (or latin square in standard form) is one in which the first row and the first column are arranged in alphabetical order, for example, for m = 3,

A B C

B C A

C A B

This is the only reduced latin square. The number of squares that can be generated from a reduced latin square by permutation of the rows, columns, and letters is (m!). These are not necessarily all different. If all rows but the first and all columns are permuted, we generate m! (m − l)! different squares. From the reduced latin square of order 3 we can thus generate 3!x(3-1)!-12squares.

In LSD two restrictions are imposed by forming blocks in two orthogonal directions, row-wise and column-wise. Further in LSD the number of treatments equals the number of replications of the treatment. Let there are m treatments and each is replicated m times then the total number of experimental units needed for the designs are m x m. These m^2 units are arranged in m rows and m columns. Then m treatments are allotted to these m^2 units at random subject to the condition that each treatment occurs once and only once in each row and in each column.

Selected Latin Squares:

3x3	**4x4**			
	1	**2**	**3**	**4**
ABC BCA CAB	ABCD BADC CDBA DCAB	ABCD BCDA CDAB DABC	ABCD BDAC CADB DCBA	ABCD BADC CDAB DCBA

5x5	6x6	7x7
ABCDE	ABCDEF	ABCDEFG
BAECD	BFDCAE	BCDEFGA
CDAEB	CDEFBA	CDEFGAB
DEBAC	DAFECB	DEFGABC
ECDBA	ECABFD	EFGABCD
	FEBADC	FGABCDE
		GABCDEF

For randomization purpose two-way heterogeneity is eliminated by means of rows and columns and a latin square of order m x mis picked up from the table of Fisher and Yates. After picking the latin square its rows and columns are randomised by the help of random numbers and this randomized square is superimposed on the arranged square.

14.3. STATISTICAL ANALYSIS OF LSD:

Let y_{ijk} (i, j, k = 1, 2, …, m) denote the response from unit (plot in the filed experimentation) in the i^th row, j^th column and receiving the k^th treatment. The triple (i, j, k) assumes only m2 of the possible m^3 values of an LSD selected by the experiment. If S represents the set of m^2 values, then symbolically (i, j, k) belongs to S. If a single observation is made per experimental unit, then the linear additive model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk};\ (i, j, k) \in S$$

where, μ is the general mean effect, α_i, β_j and τ_k are the constants effects due to the i^th row, j^th column and k^th treatment respectively and e_{ijk} is the error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2 i.e. e_{ijk} follows (i.i.d.) $N\left(0, \sigma_e^2\right)$.

If we write that

$G = y$ = Grand total of all the m^2observations.

R_i= y = Total form observations in the i^th row.

C_j= y_j = Total of the m observations in the j^th column.

T_k= $y_{..}$ = Total of the m observations in the k^{th} treatment.

Then heuristically, we get

$$\sum_i\sum_j\sum_k\left(y_{ijk}-\bar{y}...\right)^2 = \sum_i\sum_j\sum_k\left[\left(\bar{y}_{i..}-\bar{y}_{...}\right)+\left(\bar{y}_{.j.}-\bar{y}_{...}\right)+\left(\bar{y}_{..k}-\bar{y}_{...}\right)\right.$$

$$(i,j,k)\in S \qquad \left.+\left(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...}\right)\right]^2$$

$$= m\sum_i\left(\bar{y}_{i..}-\bar{y}_{...}\right)^2+m\sum_j\left(\bar{y}_{.j.}-\bar{y}_{...}\right)^2+m\sum_k\left(\bar{y}_{..k}-\bar{y}_{...}\right)^2$$

$$+\sum_i\sum_j\sum_k\left(y_{ijk}-y_{i..}-y_{.j.}-y_{..k}+2\bar{y}...\right)^2$$

The product terms vanish since the algebraic sum of deviations from mean is zero. Thus

TSS = SSR + SSC + SST + SSE

Where TSS is the total sum of squares and SSR, SSC, SST and SSE are sum of squares due to rows, columns, treatments and due to error respectively given by

$$\text{TSS} = \sum_{i,j,k\in S}\left(y_{ijk}-y_{..}.\right)^2;$$

$$\text{SSR} = m\sum_i\left(\bar{y}_{i...}-y_{...}\right)^2 = S_R^2\ (\text{say})$$

$$\text{SSC} = m\sum_J\left(\bar{y}_{.j.}-\bar{y}_{...}\right)^2 = S_C^2$$

$$= m\sum_k\left(\bar{y}_{..k}-\bar{y}_{...}\right)^2 = S_T^2$$

and SSE = S_E^2 = TSS – SSR – SSC – SST

Hence, the Total sum of squares is partitioned into three sums of squares, whose degree of freedom add to the degree of freedom of TSS.

ANOVA Table for LSD:

Source of variation	DF	SS	MSS	Variance Ratio (F)
Treatments	m-1	S_T^2	$MSST = S_T^2/(m-1)$	$F_T = \dfrac{MSST}{MSSE}$
Columns	m-1	S_C^2	$MSSC = S_C^2/(m-1)$	$F_C = \dfrac{MSSC}{MSSE}$
Rows	m-1	S_R^2	$MSSR = S_R^2/(m-1)$	$F_R = \dfrac{MSSR}{MSSE}$
Error	(m-1)(m-2)	S_E^2	$MSSE = S_E^2/(m-1)(m-2)$	
Total	m^2-1			

Under the null hypothesis,

For row effects $H_{0\alpha}$: $\alpha_1 = \alpha_2 = \ldots. = \alpha_m = 0$

For column effects $H_{0\beta}$: $\beta_1 = \beta_2 = \ldots. = \beta_m = 0$ and

For treatment effects $H_{0\tau}$: $\tau_1 = \tau_2 = \ldots. = \tau_m = 0$

against the alternative that all α's, β's and τ's are not equal, the test statistics F_T, F_c, F_R follow F distribution with [(m – 1), (m – 1) (m – 2)] df, under the above null hypothesis.

Thus, $F_\alpha = F_\alpha$ [(m– 1), (m– 1) (m– 2)] be the tabulated value of F distribution with [(m– 1), (m– 1) (m– 2)] df at the level of significance α. Thus, if $F_R > F_\alpha$ we reject the null hypothesis H0α, otherwise accept the null hypothesis. Similarly, we can test for $H_{0\beta}$ and $H_{0\tau}$.

Remark 1: Efficiency of LSD over RBD:

There may be two cases to judge the relative efficiency of LSD over RBD:

1) Relative efficiency of LSD over RBD, when rows are taken as blocks is

$$= \frac{MSSC+(m\text{-}1)MSSE}{m\,x\,MSSE}$$

2) Relative efficiency of LSD over RBD, when columns are taken as blocks is

$$= \frac{\text{MSSR}+(\text{m-1})\text{MSSE}}{\text{m x MSSE}}$$

Remark 2: Efficiency of LSD over CRD

Relative efficiency of LSD over CRD is given by

$$= \frac{\text{MSSR+MSSE}+(\text{m-1})\text{MSSE}}{(\text{m}-1)\text{MSSE}}$$

Example 1: The example of petrol consumption by different makes of cars for illustrating randomised block designs has been converted to one with 5 makes of cars to illustrate latin square design. The effects of day and driver on consumption rate have been eliminated in addition to the effect of speed by suitable modification of the experimental situation. For this purpose, 5 drivers were chosen and each driver was used on one of 5 days. On that day, he drove5 cars each of different make and each car with a different speed. The arrangement of the drivers, speeds and makes was as in the following table:

		Speeds in Miles Per Hour				
		25	**35**	**50**	**60**	**70**
Drivers and Days	**D1**	B(19.5)	E(21.7)	A(18.1)	D(14.8)	C(13.7)
	D2	D(16.2)	B(19.0)	C(16.3)	A(17.9)	E(17.5)
	D3	A(20.6)	D(16.5)	E(19.5)	C(15.2)	B(14.1)
	D4	E(22.5)	C(18.5)	D(15.7)	B(16.7)	A(16.0)
	D5	C(20.5)	A(19.5)	B(15.6)	E(18.7)	D(12.7)

Solution:

Here, Di (i = 1, 2, 3, 4, 5) denotes the ith driver driving in the ith day. A, B, C, D and E denote the 5 Makes of the cars. In the first cell of the table indicates that a car of Make B was driven by D1 on this day with a speed of 25 miles per hour. The alphabets in the other cells have similar meaning. The number of miles covered by a gallon of petrol is shown in bracket in each cell.

The design adopted is actually a latin square design with the makes of cars as treatments and the drivers and speeds are the two controlled factors representing rows and columns. The observations of the miles per hour have been analysed below as appropriate for the design.

Correction Factor = 7638.76

Sum of Squares due to Speeds = 7719.49 –7638.76 = 80.73

Sum of Squares due to Drivers = 7640.12 – 7638.76 = 1.36

Sum of Squares due to Makes = 7704.18 – 7638.76 = 65.42

Total Sum of Squares = 7792.70 – 7638.76 = 153.94

Error Sum of Squares = 153.94 – 80.73 – 1.36 – 65.42 = 6.43

ANALYSIS OF VARIANCE TABLE

Sources of Variation	DF	SS	MS	F Calculated	F Tabulated
Speeds	4	8073	20.18	37.37**	3.26
Drivers	4	1.36	0.34	0.63	
Makes	4	65.42	16.35	30.28**	
Error	12	6.43	0.54		
Total	24	153.94			
**** highly significant**					

Mean numbers of miles per gallon for the different makes arranged in order

$\bar{E}$	$\bar{A}$	$\bar{B}$	$\bar{C}$	$\bar{D}$
19.98	18.42	16.98	16.84	15.18

$$SE = \sqrt{\frac{2xMSSE}{5}} = \sqrt{\frac{2x0.54}{5}} = 0.33$$

CD at 1 per cent = 3.055 x 0.33 = 1.42

The initial difference indicates that the Make E is significantly better than all the other Makes. Make A was better than B, C and D. Finally, D is the worst.

Efficiency of Latin square

$$\text{E (Drivers)} = \frac{4\text{x}0.34\text{+}0.54\text{x}16}{20\text{x}0.54} = \frac{34 + 0.54\text{x}4}{5\text{x}0.54} = 0.93$$

$$\text{E (Speeds)} = \frac{4\text{x}20.18\text{+}0.54\text{x}16}{20\text{x}0.54} = \frac{20.18 + 0.54\text{x}4}{5\text{x}0.54} = 8.27$$

The efficiency figures show that elimination of speed variation increased precision considerably while elimination of driver variation did not reduce error variance.

14.4. MISSING PLOTS TECHNIQUE IN LSD:

As we have discussed in Section 10.4 of Unit 10, sometimes observations from one or more experimental units are not found (missing) due to some unavoidable causes. There may be some unforeseen causes for example in agricultural experiments damage by animal or pets, in animal experiment any animal may die or observations from one or more plot is excessively large as compared to other plots and thus accuracy of such observation is often in doubt. In such situations, these observations are omitted and treated as missing.

In case of missing observations, analysis is done by estimating the missing observation. This type of analysis was given by Yates (1937) and it is known as missing plot technique. As similar as in the RBD, we are now going to discuss the same in LSD in the following sub-section:

14.4.1. One Missing Plot:

Suppose without loss of generality that in m x m latin square design the observation occurring in the first row, first column and receiving first treatment is missing. Let us assume that y_{111}=Y

R'_1 = Total of all available (m – 1) observations in 1st row.

C'_1 = Total of all available (m – 1) observations in 1st column.

T'_1 = Total of all available (m – 1) observations receiving 1st treatment

G' = Total of all available ($m^2 - 1$) observations.

On the basis of these totals we calculate different sum of squares as follows:

$$\text{Sum of Squares for Rows (SSR)} = \frac{\left(R_1^{'} + Y\right)^2 + \sum_{i=2}^{m} R_i^2}{m} - \frac{\left(R_1^{'} + Y\right)^2}{m^2}$$

Sum of Squares for Columns (SSC) = $\dfrac{\left(C_1^{'}+Y\right)^2+\sum\limits_{j=2}^{m}C_j^2}{m}-\dfrac{\left(G^{'}+Y\right)^2}{m^2}$

Sum of Squares for Treatments (SST) = $\dfrac{\left(T_1+Y\right)^2+\sum\limits_{k=2}^{m}T_K^2}{m}-\dfrac{\left(G^{'}+Y\right)^2}{m^2}$

Total Sum of Squares (TSS) = $\sum\limits_{i}\sum\limits_{j}\sum\limits_{k}y_{ijk}^2+Y^2-\dfrac{\left(G'+Y\right)^2}{m^2}$

$$(i,j,k)\neq(1,1,1)$$

Sum of Squares due to Error (SSE) = TSS – SSR – SSC – SST

$$SSE = Y^2+\frac{2\left(G'Y\right)^2}{m^2}-\frac{2\left(R'_1Y\right)^2}{m}$$

$$-\frac{\left(C_1^{'}+Y\right)^2}{m}-\frac{\left(T_1^{'}+Y\right)^2}{m}\text{Terms not involving Y}$$

For obtaining the value of Y, we minimize the sum of squares due to error with respect to Y. This is obtained by solving the equation

$$\frac{\partial(SSE)}{\partial Y}=2Y+\frac{4\left(G'+Y\right)}{m^2}-\frac{2\left(R_1^{'}+Y\right)}{m}-\frac{2\left(C_1^{'}+Y\right)}{m}-\frac{2\left(T_1^{'}+Y\right)}{m}=0$$

$$\Rightarrow Y+\frac{2Y}{m^2}-\frac{Y}{m}-\frac{Y}{m}-\frac{Y}{m}=\frac{R_1^{'}}{m}+\frac{C_1^{'}}{m}+\frac{T_1^{'}}{m}-\frac{2G^{'}}{m^2}$$

$$\Rightarrow\frac{Y\left(m^2+2-3m\right)}{m^2}=\frac{m\left(R_1^{'}+C_1^{'}+T_1^{'}\right)-2G'}{m^2}$$

$$\hat{Y}=\frac{m\left(R_1^{'}+C_1^{'}+T_1^{'}\right)-2G'}{(m-1)(m-2)}$$

$\hat{Y}$ is the least square estimate of the yield of the missing plot. The value of Y is inserted in the original table of yield and ANOVA is performed in the usual way except that for each missing observation 1 df is subtracted from total and consequently from error df.

Example 2:

In the following data, one value is missing. Estimate this value and analyse the given data.

Column **Row**	**I**	**II**	**III**	**IV**	**Row Totals (R1)**
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D Y	A 5	C 21	48+Y
IV	D 12	A 7	C 27	B 17	63
Column Totals (Cj)	64	38+Y	48	53	203+Y

Solution: Here $m = 4, R_3^{'} = 48, C_2^{'} = 38,\ T_4^{'} = 26,\ G^{'} = 203$

Applying the missing estimation formula

$$\hat{Y} = \frac{m\left(R_3^{'} + C_2^{'} + T_4^{'}\right) - 2G'}{(m-1)(m-2)}$$

$$= \frac{4(48+38+26) - 2x203}{(4-1)(4-2)}x7$$

Inserting the estimated value of Y, we get the following observations:

Column **Row**	**I**	**II**	**III**	**IV**	**Row Totals (R₁)**
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D Y	A 5	C 21	48+Y
IV	D 12	A 7	C 27	B 17	63
Column Totals (C_j)	64	45	48	53	203

Correction Factor (CF) $= \dfrac{(210)^2}{16} = \dfrac{44100}{16} = 2756.25$

Raw Sum of Squares (RSS) = $(12)^2+(18)^2+ \ldots + (21)^2+ (17)^2 = 3432$

Total Sum of Squares (TSS) = 3432– 2756.25 = 675.75

Row Sum of Squares (SSR) $= \dfrac{(49)^2+(43)^2+(55)^2+(63)^2}{4} - \text{CF}$

$$= \frac{2401+1849+3025+3969}{4} - 2756.25 = 54.75$$

Column Sum of Squares (SSC) $= \dfrac{(64)^2+(45)^2+(48)^2+(53)^2}{4} - \text{CF}$

$$= \frac{4096+2025+2304+2809}{4} - 2756.25 = 52.25$$

Treatment Sum of Squares (TSS) $= \dfrac{(31)^2 + (61)^2 + (85)^2 + (33)^2}{4} - \text{CF}$

$$= \frac{961 + 3421 + 7225 + 1089}{4} - 2756.25 = 417.75$$

Error Sum of Squares (SSE) = TSS– SSR – SSC – SST

=675.75–54.75–52.25–417.75=151

ANOVA TABLE

Source of Variation	DF	SS	MSS	Variance Ratio		Conclusion
				Calculated	Tabulated	
Rows	4– 1=3	54.75	18.25	0.60	5.41	Insignificant
Columns	4– 1=3	52.25	17.42	0.58	5.41	Insignificant
Treatments	4– 1=3	417.75	139.25	4.61	5.41	Insignificant
Error	6– 1=5	151	30.20			
Total	15– 1 =14					

14.5. SUITABILITY OF LSD:

The latin square design is used when the experimental material is heterogeneous with respect to two factors and this two-way heterogeneity is eliminated by means of rows and columns. In fact, LSD can be applied to all those cases where either the variation in the experimental material is not known or is known in two mutually perpendicular directions. Thus, LSD is successfully used in industry, animal husbandry, biological and social sciences, piggeries, marketing, medical and educational fields, where it is desired to eliminate the two-factor heterogeneity simultaneously.

Advantages and Disadvantages of LSD:

Advantages of LSD:

1) Since total variation is divided into three parts namely rows, columns and treatments, the error variance is reduced considerably. It happens due to the fact that rows and columns being perpendicular to each other, eliminates the two-way heterogeneity up to a maximum extent.

2) LSD is an incomplete three-way layout. Its advantage over the complete three-way layout is that instead of m^3 units only m^2 units are needed. Thus, a 4x4 LSD results in saving 64 – 16 = 48 observations over a complete three-way layout.

3) The analysis creates no problem even if a missing observation exists.

Disadvantages of LSD:

1) The fundamental assumption that there is no interaction between different factors may not be true in general.

2) The main limitation of LSD is the equality of number of rows to that of columns and treatments. If the layout of experimental material is not of square design then LSD cannot be used.

3) RBD can be accommodated in any shape of field whereas for LSD field should perfectly be a square.

4) For smaller number of treatments, say less than 5, the degree of freedom for error is very small and thus the results are not reliable. Even in case of2x2 LSD, degree of freedom for error becomes zero. In such situations,either the number of treatments should be increased or the latin square should be repeated.

5) On the other side, if the number of treatments increases the size of latin squares increases and this causes a disturbance in heterogeneit y.

6) Analysis of LSD becomes very much complicated if complete row or complete column is missing. Analysis of RBD is quite easy in such situations.

14.6. SUMMARY:

In this Unit, we have discussed:

1) The Latin Square design;

2) The layout of LSD;

3) The method of statistical analysis of LSD;

4) The missing plots technique in LSD; and

5) The advantages and disadvantages of LSD.

Example 3: Carry out ANOVA for the following design:

A **5**	**B** **7**	**C** **7**	**D** **8**	**E** **9**
B **7**	**C** **9**	**D** **8**	**E** **8**	**A** **5**
C **6**	**D** **5**	**E** **9**	**A** **8**	**B** **9**
D **5**	**E** **6**	**A** **8**	**B** **5**	**C** **7**
E **8**	**A** **9**	**B** **5**	**C** **7**	**D** **6**

The analysis of the given design is done by the method of analysis of variance. The computation results are given as follows:

Correction factor (CF)	=	1239.04
Raw Sum of Squares	=	1292
Total Sum of Squares	=	52.92
Column Sum of Squares	=	4.56
Row Sum of Squares	=	4.96
Treatment Sum of Squares	=	7.76
Error Sum of Squares	=	35.68

ANALYSIS OF VARIANCE TABLE

Sources of Variation	**DF**	**SS**	**MSS**	**F**
Rows	4	4.96	1.24	0.42
Columns	4	4.56	1.14	0.38
Treatment	4	7.76	1.94	0.65
Error	12	35.68	2.97	
Total	24	52.96		

Tabulated value of F (4, 12) = 3.26

Since the calculated value of F is much less than the tabulated value of F at 5% level of significance, we conclude that there is no significant difference between treatment means.

Example 4: Let the missing value is Y then we have

Column **Row**	**I**	**II**	**III**	**IV**	**Row Totals (Ri)**
I	A 8	C 18	B 11	D 8	45
II	C 16	B 10	D 7	A Y	33+Y
III	B 12	D 10	A 6	C 20	48
IV	D 10	A 9	C 28	B 16	63
Column **Totals (Cj)**	46	47	52	44+Y	189+Y

Here, $m = 4,\ R_2^{'} = 33, C_4^{'} = 44, T_1^{'} = 23,\ G' = 189$

Applying the missing estimation formula

$$\hat{Y} = \frac{m\left(R_3^{'} + C_2^{'} + T_4^{'}\right) - 2G'}{(m-1)(m-2)}$$

$$= \frac{4(33+44+23) - 2x189}{(4-1)(4-2)} = 3.66 \sim 4$$

Inserting the estimated value of Y, we get the following observations:

Column **Row**	**I**	**II**	**III**	**IV**	**Row Totals (Ri)**
I	A 8	C 18	B 11	D 8	45
II	C 16	B 10	D 7	A 4	37
III	B 12	D 10	A 6	C 20	48
IV	D 10	A 9	C 28	B 16	63
Column Totals (Cj)	46	47	52	48	193

Correction Factor (CF) $= \frac{(193)^2}{16} = 2328.06$

Raw Sum of Squares (RSS) = = $(8)^2$ x $(16)^2$ x ... x $(20)^2$ x $(16)^2$= 2895

Total Sum of Squares (TSS) = = 2895 – 2328.06 = 566.94

Row Sum of Squares (SSR) $= \frac{(45)^2 + (37)^2 + (48)^2 + (63)^2}{4} - CF$

$$= \frac{2025+1369+2304+3959}{4} - 2328.06 - 88.69$$

$$\text{Column Sum of Squares (SSC)} = \frac{(46)^2+(47)^2+(52)^2+(48)^2}{4} - \text{CF}$$

$$= \frac{2116+2209+2704+2304}{4} - 2328.06 = 5.19$$

$$\text{Treatment Sum of Squares (SST)} = \frac{(27)^2+(49)^2+(82)^2+(35)^2}{4} - \text{CF}$$

$$= \frac{729+2401+6724+1225}{4} - 2328.06 = 441.69$$

Error Sum of Squares (SSE) = TSS– SSR – SSC – SST

= 566.94 – 88.69 – 5.19 – 441.69 = 31.37

ANOVA TABLE

Source of Variation	**DF**	**SS**	**MSS**	**Variance Ratio**		**Conclusion**
				Calculated	**Tabulated**	
Columns	4−1=3	5.19	1.73	0.28	5.41	Insignificant
Treatments	4−1=3	441.69	147.23	23.48	5.41	Significant
Error	6−1=5	31.37	6.27			
Total	15−1=14					

Since for treatment effect calculated value of F is greater than the tabulated value of F at 5% level of significance, so we conclude that the treatment effect is significant. For pairwise testing, find the standard error of difference of two treatment means.

$$SE = \sqrt{\frac{2MSSE}{m}} = \sqrt{\frac{2x6.27}{4}} = 1.77$$

Critical difference (CD) = $SE\ x\ t_{\alpha/2}$ at error df

$$= 1.77 \text{ x } 2.571 = 4.55$$

Treatment means

$$\overline{A} = \frac{27}{4} = 6.75,\ \overline{B} = \frac{49}{4} = 12.25,\ \overline{C} = \frac{82}{4} = 20.5\ \&\ \overline{D} = \frac{35}{4} = 8.75$$

Pair of Treatments	**Difference**	**CD**	**Inference**
A, B	$\left\|\overline{A} - \overline{B}\right\| = 05.50$	4.55	Significant
A, C	$\left\|\overline{A} - \overline{C}\right\| = 13.75$	4.55	Significant
A, D	$\left\|\overline{A} - \overline{D}\right\| = 02.00$	4.55	Insignificant
B, C	$\left\|\overline{B} - \overline{C}\right\| = 08.25$	4.55	Significant
B, D	$\left\|\overline{B} - \overline{D}\right\| = 03.50$	4.55	Insignificant
C, D	$\left\|\overline{C} - \overline{D}\right\| = 11.75$	4.55	Insignificant

14.7. SELF-ASSESSMENT QUESTIONS:

1) Explain Layout of Latin Square Design
2) Explain the Statistical Analysis of Latin Square Design
3) Explain the Least Square Estimates of Effects of LSD
4) Missing Plots Technique in LSD
5) Explain the Suitability of LSD

14.8. SUGGESTED READINGS:

1) Montgomery, D.C. (2017). *Design and Analysis of Experiments* (9th ed.). Wiley.

2) Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

3) Gupta, S.C., & Kapoor, V.K. (2014). *Fundamentals of Applied Statistics*. Sultan Chand & Sons.

Dr. B. Hari Mallikarjuna Reddy

LESSON-15

TEST OF NORMALITY

15.0. OBJECTIVES:

After completing this unit, you should be able to:

- Understand the importance of the normality assumption in linear models and ANOVA.
- Explain the theoretical basis of normality in residuals.
- Apply graphical and numerical tests for normality.
- Interpret the results of Shapiro–Wilk, Kolmogorov–Smirnov, Anderson–Darling, and Chi-square tests.
- Recognize limitations and practical considerations in testing normality.

STRUCTURE:

15.1. INTRODUCTION:

Statistical models such as linear regression and Analysis of Variance (ANOVA) rely on certain assumptions. One of the most fundamental assumptions is that the error terms (residuals) follow a normal distribution with mean zero and constant variance. This assumption simplifies inference because many test statistics (t, F, χ^2) are derived under normality. Normality refers to the condition where a dataset follows a Normal Distribution (also called the Gaussian distribution or bell curve). The normal distribution is symmetrical, with most of the data clustered around the mean and tapering off equally on both sides.

It is Fully Described by Two Parameters:

- Mean (μ) → the central tendency
- Variance (σ^2) → the spread of data

Normality Ensures:

- Reliable estimation of parameters.
- Valid hypothesis testing.
- Correct confidence intervals.
- Without normality, results may be misleading, especially in small samples.

15.2. ROLE OF NORMALITY IN LINEAR MODELS AND ANOVA:

- In ANOVA, the F-statistic assumes residuals are normally distributed.
- In linear regression, least squares estimate remain unbiased without normality, but t-tests and F-tests may be invalid.
- The Central Limit Theorem (CLT) suggests that with large samples, normality is less critical; but with small or moderate samples, normality should be checked.

Testing Normality is Important because:

1. **Statistical Assumptions**
   - Many parametric tests (t-test, ANOVA, regression) assume that the data (or residuals) follow a normal distribution.
   - If the assumption is violated, results may be misleading.
2. **Model Adequacy Checking**
   - In regression and ANOVA, residuals should be normally distributed.
   - Non-normality may suggest model misspecification.
3. **Practical Decision Making**
   - Helps decide whether to use parametric methods (require normality) or non-parametric methods (do not assume normality).

15.3. CONSEQUENCES OF NON-NORMALITY:

Type I error inflation – p-values may be inaccurate.

Loss of power – tests may fail to detect significant differences.

Biased parameter estimates when distributions are skewed or have outliers.

Heteroscedasticity Can Worsen Violations – unequal variances across groups make normality departures more problematic.

Outliers Strongly Affect Normality Tests – even a single extreme value can cause the test to reject normality.

Tests Rely on the Assumption of Continuous Data – discretized or rounded data can lead to misleading conclusions.

15.4. GRAPHICAL METHODS FOR CHECKING NORMALITY:

These give a visual check of distribution:

- Histogram: Compare shape to bell curve.
- Boxplot: Detect skewness and outliers.
- Q-Q Plot (Quantile-Quantile Plot): If points lie close to a straight diagonal line, data is approximately normal.
- P-P Plot (Probability-Probability Plot): Compares cumulative probabilities.
- Histogram of Residuals – should resemble bell-shaped curve if normal.

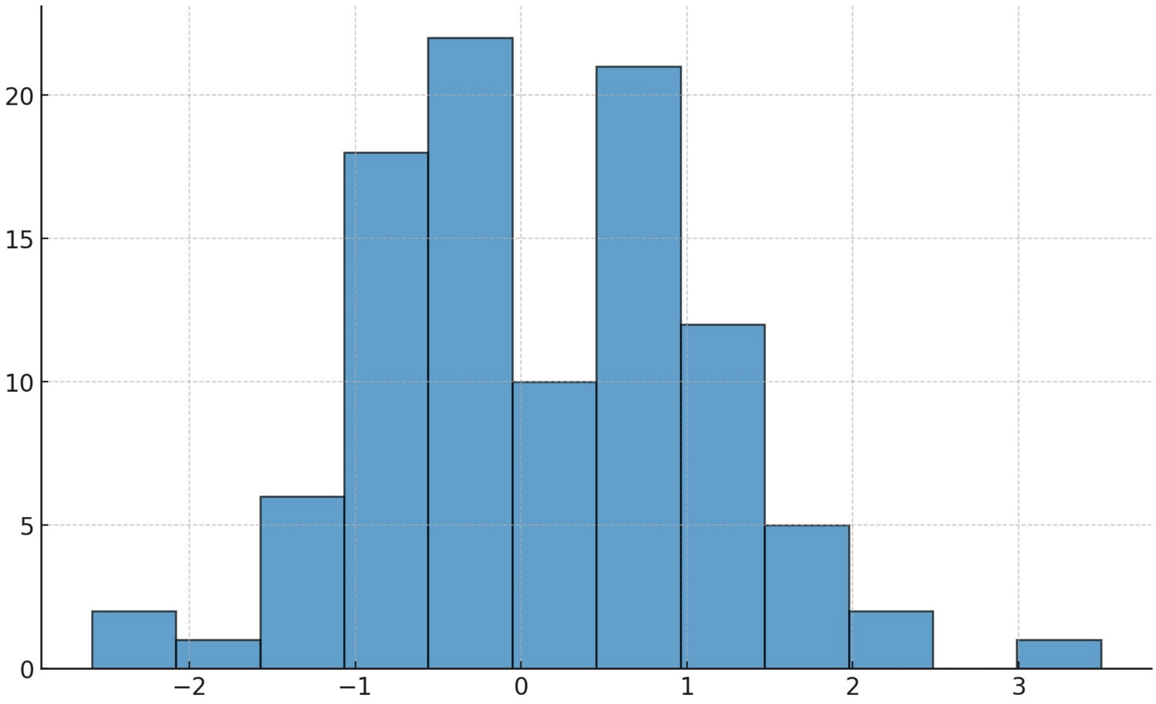

Figure 15.1: Histogram of Normally Distributed Residuals

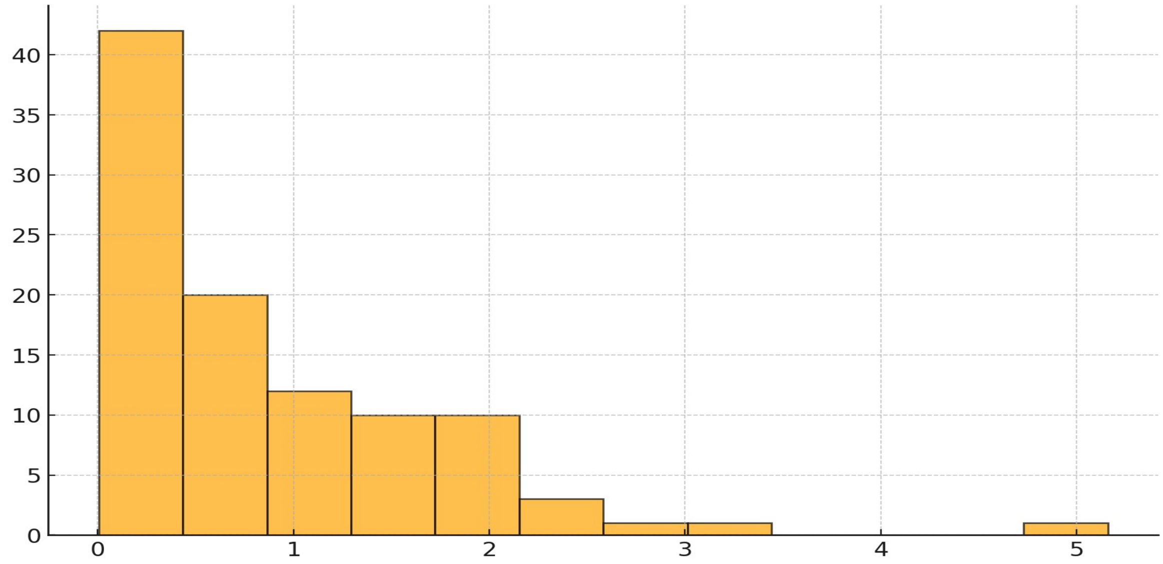

Figure 15.2: Histogram of Skewed Residuals

Box Plot – Detects Skewness and Outliers

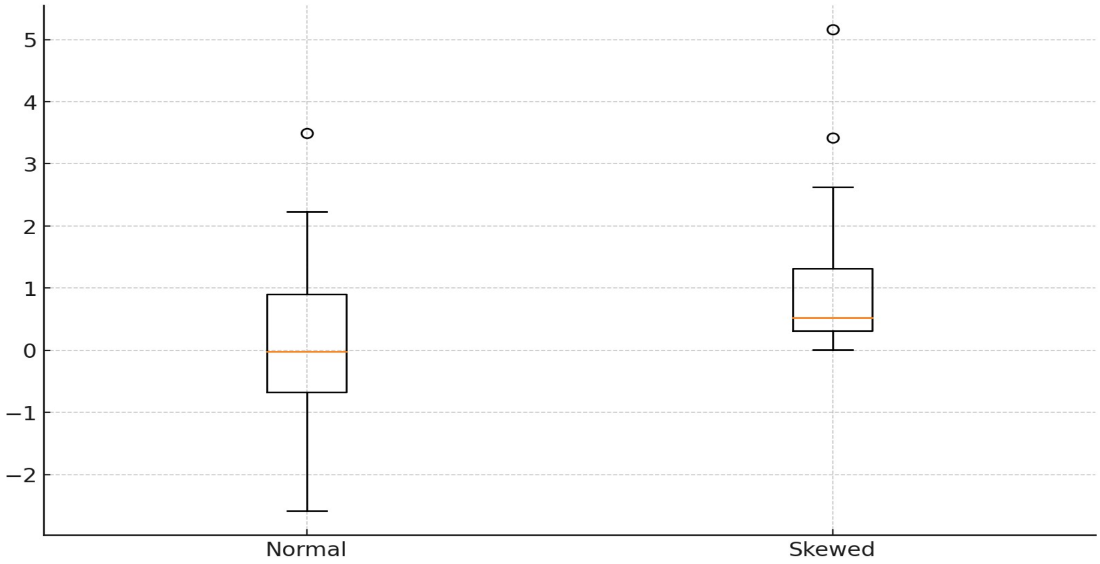

Figure 15.3: Boxplot Comparison (Normal vs Skewed)

Q-Q Plot (Quantile–Quantile Plot) – Observed Quantiles vs Theoretical Quantiles; should lie on 45° line.

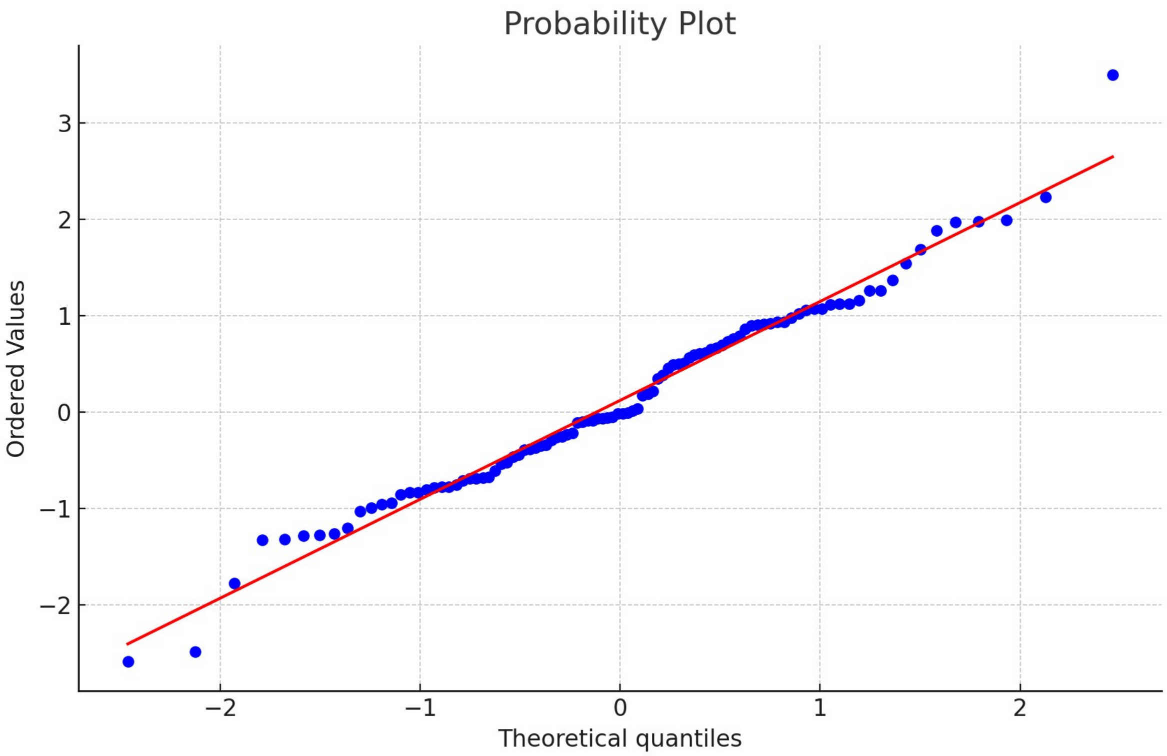

Figure 15.4: Q-Q Plot for Normal Data

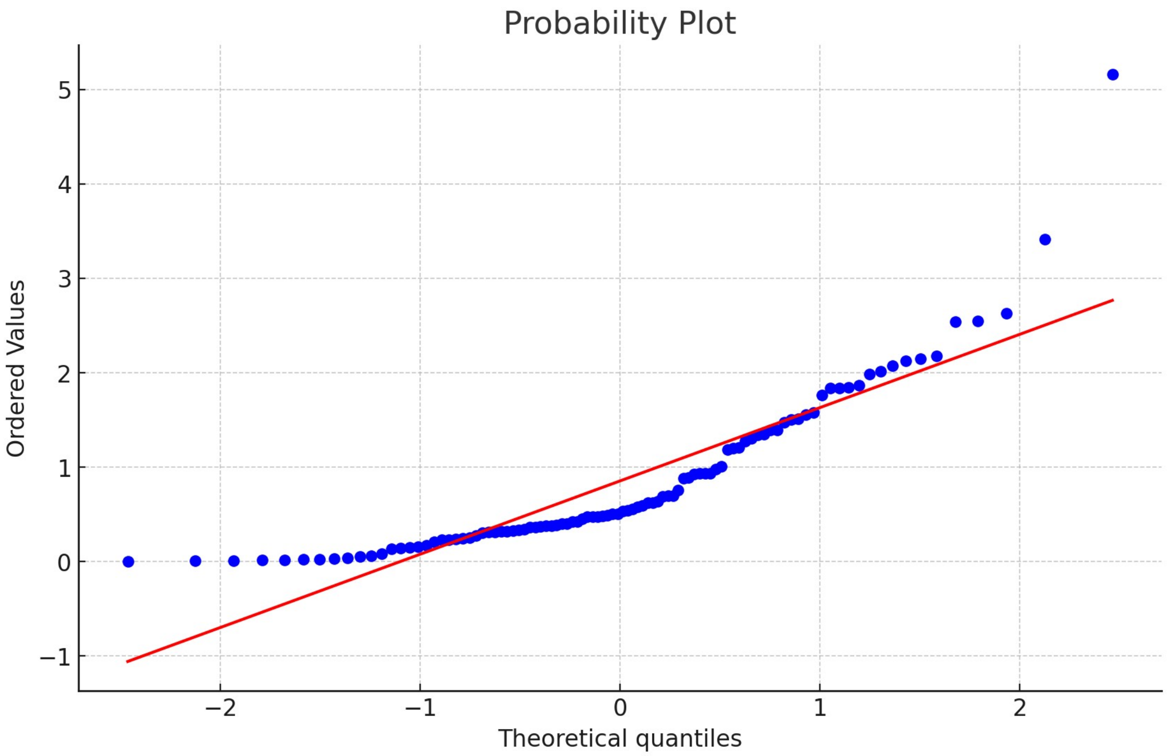

Figure 15.5: Q-Q Plot for Skewed Data

P-P Plot (Probability–Probability Plot) – Plots Cumulative Probabilities.

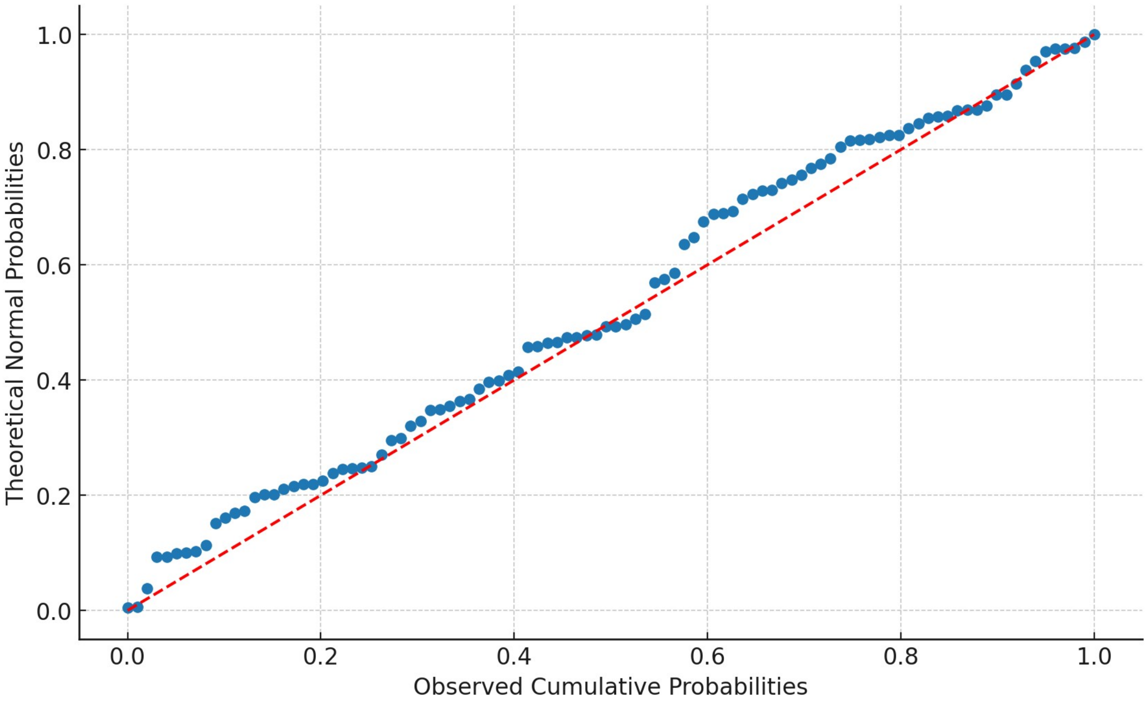

Figure 15.6: P-P Plot for Normal Data

15.5. STATISTICAL TESTS FOR NORMALITY:

15.5.1. Shapiro–Wilk Test:

- The Shapiro–Wilk test is a statistical test for normality.
- It checks whether a sample comes from a normally distributed population.
- First proposed by Shapiro and Wilk (1965).
- It is one of the most powerful and widely used normality tests, especially for small to medium sample sizes.

Hypotheses:

- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_1): The data are not normally distributed.

Statistic (W): The test calculates a statistic W defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where,

x_i = the ordered sample values (from smallest to largest).

$\bar{x}$= Sample Mean

a_i= constants derived from the covariance matrix of the order statistics of a normal distribution.

W ranges between 0 and 1. Values closer to 1 indicate stronger normality.

Decision Rule:

- A p-value is computed from W.
- If $p > 0.05$ → Fail to reject H_0 → Data is approximately normal.
- If $p < 0.05$ → Reject H_0 → Data significantly deviates from normality.

Data (exam scores): 52, 55, 60, 62, 64, 65, 66, 68, 70, 72

We test

- H_0: data are from a normal distribution
- H_1: data are not from a normal distribution

Step 1: Order the sample:data are not from a normal distribution

$x_{(1)} \le x_{(2)} \le \cdots \le x_{(10)}$

Already sorted:

$$x_{(1)} = 52,\ x_{(2)} = 55,\ x_{(3)} = 60,\ x_{(4)} = 62,\ x_{(5)} = 64,\ x_{(6)} = 65,\ x_{(7)} = 66,\ x_{(8)} = 68,\ x_{(9)} = 70,\ x_{(10)} = 72$$

Step 2: Compute the sample mean and the total sum of squares

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{52 + 55 + \cdots + 72}{10} = 63.4$$

$$\text{SS} = \sum (x_i - \bar{x})^2 = (52 - 63.4)^2 + \cdots + (72 - 63.4)^2 = 362.4$$

This SS is the denominator of Shapiro–Wilk's W.

Step 3: Get the Shapiro–Wilk weights

The test forms a symmetric, weighted contrast of the order statistics:

The test forms a symmetric, weighted contrast of the order statistics:

$b\ =\ \sum_{i=1}^{n} a_i\ x_{(i)}$ (with $a_i = a_{n+1-i}$)he weights a_i are functions of the expected normal order statistics and their covariance matrix; in practice they are looked up from published tables or computed by software, not hand-derived.

For n=10, we use software (equivalent to table values) to obtain the a_i and compute b.

Step 4: Compute the Shapiro–Wilk statistic

ompute the Shapiro – Wilk statistic

$$\mathrm{W}\ =\ \frac{b^{\,2}}{\mathrm{SS}}$$

Using standard software for this dataset:

- W=0.9603
- p = 0.7890

(These reflect the exact a_i and the Monte-Carlo–based p-value approximation used by modern implementations.)

$SS\ =\ 362.4, W\ = 0.9603,\ b\ = \sqrt{W \cdot SS} = \sqrt{0.9603 \times 362.4} \approx 18.655$

15.5.2. Kolmogorov–Smirnov Test:

The K-S test is a non-parametric test used to check whether:

1) A sample comes from a specified distribution (e.g., normal distribution) → *One-sample K-S test.*
2) Two samples come from the same distribution → *Two-sample K-S test.*

It compares the empirical distribution function (EDF) of the sample with the theoretical CDF (one-sample) or compares the EDFs of two samples (two-sample).

Test Statistic (D):

$\mathrm{D} = \sup|F_{\mathrm{n}}(x) - F^{0}(x)|$

1) One-Sample K-S Test Numerical Example

Problem: We have a sample of n = 8 observations: 2,4,6,8,10,12,14,16

We want to test if the data comes from a **Uniform (0, 20)** distribution at the **5% level of significance**.

Step 1: Order the data Already ordered: 2,4,6,8,10,12,14,16

Step 2: Compute ECDF $\left(F_n(x)\right)$ $F_n(xi) = \frac{i}{n}, i = 1,2, \ldots, n$

x_i	Rank (i)	$F_n(x_i)$
2	1	$1/8 = 0.125$
4	2	$2/8 = 0.250$
6	3	$3/8 = 0.375$
8	4	$4/8 = 0.500$
10	5	$5/8 = 0.625$
12	6	$6/8 = 0.750$
14	7	$7/8 = 0.875$
16	8	$8/8 = 1.000$

Step 3: Theoretical CDF F(x)

For Uniform (0,20): $\mathrm{F}(x) = \frac{x-0}{20}, 0 \leq x \leq 20$

x	$\mathrm{F}(xi)$
2	$2/20 = 0.10$
4	$4/20 = 0.20$
6	$6/20 = 0.30$
8	$8/20 = 0.40$
10	$10/20 = 0.50$
12	$12/20 = 0.60$
14	$14/20 = 0.70$
16	$16/20 = 0.80$

We calculate $D^+ = \max\left(F_n(x_i) - F(x_i)\right)$ and $D^- = \max\left(F(x_i) - F_n(x_{i-1})\right)$

- Test statistic: $\mathrm{D} = \max(D^+, D^-)$

x_i	$F_n(x_i)$	$\mathrm{F}(x_i)$	$\lvert F_n - F\rvert$
2	0.125	0.10	0.025
4	0.250	0.20	0.050
6	0.375	0.30	0.075
8	0.500	0.40	0.100
10	0.625	0.50	0.125
12	0.750	0.60	0.150
14	0.875	0.70	0.175
16	1.000	0.80	0.200

So, $\mathrm{D} = \max\lvert F_n(x_i) - F(x_i)\rvert = 0.200$

Step 5: Critical Value

For, $\mathrm{n} = 8,$ significance level $\alpha = 0.05$:

$$D_{critical} = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{8}} \approx 0.48$$

We fail to reject H_0.

Conclusion: The sample is consistent with a Uniform (0,20) distribution.

15.5.3. Anderson–Darling Test:

$$A^2 = -\mathrm{n} - (1/n)\Sigma\left[(2i-1)lnF\left(x_{(i)}\right) + (2n+1-2i)ln\left(1 - F\left(x_{(i)}\right)\right)\right]$$

Example: Residuals gave $A^2 = 0.48 <$ critical $0.75 \Rightarrow$ Normal.

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

The Chi-Square (χ^2) Test is a statistical hypothesis test that compares observed data with expected data according to some assumption. It helps us check whether differences between observed and expected values are due to chance or statistically significant.

Types of Chi-Square Tests:

There are mainly two types:

1. **Chi-Square Goodness-of-Fit Test**
   - Checks if a sample data fits a particular theoretical distribution.
   - Example: Testing if a die is fair (uniform distribution).
2. **Chi-Square Test of Independence**
   - Checks if two categorical variables are independent.
   - Example: Testing if gender is related to preference for a product.

Test Statistic Formula

The general form of the chi-square statistic is:

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

Where:

- $O_i = Observedfrequency$
- E_i = Expected frequency
- The sum is taken over all categories/cells

Degrees of Freedom (df)

- For Goodness of Fit: $\mathrm{df} = \mathrm{k} - 1 - \mathrm{m}$

(where k = number of categories, m = parameters estimated)

- For Independence Test: $\mathrm{df} = (r - 1)(c - 1)$

  $(where\ r = rows, c = columns\ in\ a\ contingency\ table)$

Decision Rule

Compare calculated χ^2 value with the critical value from the chi square distribution table at a chosen significance level(α).

If $\chi^2_{calculated} > \chi^2_{critical}, reject H_0. Otherwise, failtoreject H_0.$

ASSUMPTIONS:

1. **Independence of Observations:**
   - The data points must be independent; correlation or repeated measurements violate this assumption.
2. **Continuous Scale of Measurement:**
   - Normality tests assume data are measured on a continuous scale, not categorical or overly discretized.
3. **Random Sampling**
   - The sample should be randomly drawn from the population to ensure valid inference.
4. **No Extreme Outliers**
   - Many normality tests (especially Shapiro–Wilk) are highly sensitive to outliers, which can distort results.
5. **Underlying Distribution must be Fully Specified**
   - Tests like Kolmogorov–Smirnov require that the expected distribution (e.g., N (μ, σ^2)) is known.
6. **Sufficient Sample Size**
   - Very small samples may fail to detect non-normality; very large samples may detect trivial deviations.
7. **Residual-Based Testing in Models**
   - In ANOVA or regression, normality should be checked on residuals, not raw data.

15.6. LIMITATIONS OF NORMALITY TESTS:

- Large samples may reject normality for trivial deviations.
- Small samples may fail to detect non-normality.
- Always combine graphical + statistical methods.
- Apply transformations (log, square root, Box-Cox).
- **Different tests have different sensitivities** (e.g., Shapiro-Wilk detects tail issues better than Kolmogorov-Smirnov).

- **Tests assume independent observations**; correlated data (time series, repeated measures) can give misleading results.

- **Non-normality of the data vs. non-normality of residuals** is often confused-tests should be applied to residuals in regression/ANOVA, not to raw data.

15.7. SUMMARY:

Normality is one of the fundamental assumptions underlying **ANOVA** and **linear regression**, because these methods rely on the idea that the **errors (residuals)** of the model follow a normal distribution. When this assumption holds, the resulting test statistics (such as the F-statistic and t-statistic) follow their theoretical distributions, which ensures that p-values, confidence intervals, and hypothesis tests are valid and reliable. If the residuals deviate significantly from normality, the Type I error rate may increase, estimates may become biased, and conclusions drawn from the model may be misleading. Therefore, assessing normality is an essential diagnostic step before interpreting ANOVA or regression results.

To evaluate normality, both **graphical** and **statistical** methods should be applied because each provides complementary information. Graphical tools-such as **histograms**, **Q-Q plots**, **boxplots**, and **residual plots**-help visualize the shape, symmetry, skewness, and presence of outliers. Statistical tests, such as **Shapiro–Wilk**, **Kolmogorov–Smirnov**, **Anderson–Darling**, and **Jarque-Bera**, provide quantitative evidence about deviations from normality. Among these, the Shapiro–Wilk test is widely recommended, especially for **small to medium sample sizes (n < 50 or n < 200)**, because it has high power in detecting non-normal patterns.

Normality tests should not be viewed as rigid pass–fail tools; instead, they guide the analyst in evaluating whether the assumptions behind ANOVA and regression are sufficiently met for valid inference. Even if slight departures from normality are detected, ANOVA and regression are generally robust-particularly with larger sample sizes-due to the Central Limit Theorem. However, severe deviations may require corrective actions, such as transforming the data (log, square-root, Box–Cox), using non-parametric alternatives, or applying robust statistical methods. Thus, normality assessment plays a crucial role in ensuring the accuracy, reliability, and interpretability of statistical modelling results.

15.8. SELF-ASSESSMENT QUESTIONS:

1) Why is normality of residuals required in ANOVA?

2) Compare Q-Q plots and histograms in testing normality.

3) Derive the test statistic for the Shapiro–Wilk test.

4) Why is Anderson–Darling more sensitive in the tails?

5) Discuss limitations of the Chi-square test for normality.

15.9. SUGGESTED READINGS:

1) Kutner, Nachtsheim, Neter and Li - Applied Linear Statistical Models

2) Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining - Introduction to Linear Regression Analysis

3) N.R. Draper and H. Smith - Applied Regression Analysis

4) Samprit Chatterjee and Ali S. Hadi - Regression Analysis by Example

5) S.C. Gupta and V.K. Kapoor - Fundamentals of Mathematical Statistics.

Dr. M. Amulya

LESSON-16

TEST OF EQUALITY OF VARIANCES

16.0. OBJECTIVES:

After completing this unit, you should be able to:

- Understand the assumption of **homogeneity of variances** (homoscedasticity) in ANOVA and regression.
- Explain why variance equality is essential for valid F-tests.
- Apply **Bartlett's test** and the **Modified Levene test**.
- Interpret test results and handle violations of variance homogeneity.

STRUCTURE:

16.1. INTRODUCTION:

ANOVA assumes that the populations being compared have the same variance (σ^2). This assumption is called **homogeneity of variances**. For example, if two fertilizers are compared on crop yields but one group has much larger variability than the other, the ANOVA test may show a significant difference that is actually due to unequal variability rather than a true treatment effect.

- When variances are equal, the pooled error term (MSE) provides a valid estimate of error variance.
- When variances are unequal (heteroscedasticity), the F-distribution used in ANOVA no longer holds true.
- This leads to incorrect conclusions about treatment effects.

16.2. IMPORTANCE OF EQUAL VARIANCES:

Test	Assumptions	Robustness	Best Use Case
Bartlett's Test	Assumes **normality**.	Very sensitive to non-normality & outliers.	When data is **strictly normal** and sensitivity is needed.
Levene's Test (Mean version)	Fewer assumptions; not strict normality.	Moderately robust.	General-purpose, when normality is doubtful.
Levene's Test (Modified / Brown–Forsythe, Median version)	Uses median instead of mean.	Highly robust to **non-normality & outliers**.	Most recommended in real-world data.
O'Brien's Test	Assumes approximate normality.	More powerful than Levene in some cases.	When mild deviations from normality are expected.
Hartley's F-Max Test	Assumes normality and equal sample sizes.	Very sensitive, limited use.	Quick check with small, normal samples.
Fligner–Killeen Test	Non-parametric. No normality assumption.	Very robust, distribution-free.	When data is **heavily non-normal or ordinal**.

16.3. CONSEQUENCES OF HETEROSCEDASTICITY:

- **Bias in test results**: Groups with smaller variances appear more "stable" and inflate F-ratio.
- **Loss of statistical power**: Large differences in variance reduce sensitivity of tests.
- **Misleading confidence intervals**: Standard errors are underestimated or overestimated.

Example:

In a clinical trial, if variability in the placebo group is much larger than in treatment groups, the ANOVA may incorrectly detect differences.

16.4. GRAPHICAL METHODS FOR CHECKING HOMOGENEITY:

Before applying formal tests, **visual checks** are recommended:

1) **Residual Plots**
   - Plot residuals vs fitted values.
   - If spread is constant across all fitted values ⇒ homogeneity.
   - Funnel-shaped patterns ⇒ heteroscedasticity.
2) **Box Plots**
   - Compare spread of groups visually.
   - Unequal lengths of boxes/whiskers suggest unequal variance.
3) **Spread-Level Plot**
   - Plots spread against mean.
   - Helps decide if a transformation (log, square root) is needed.

16.5. BARTLETT'S TEST OF HOMOGENEITY:

- **Bartlett's Test** is a statistical test used to check whether multiple samples (from different groups) have **equal variances**.
- It is commonly used as an **assumption check before performing ANOVA (Analysis of Variance)** because ANOVA assumes homogeneity of variances.
- The test statistic follows a **Chi-square (χ²) distribution**.
- $H_0 : \sigma^{12} = \sigma^{22} = \cdots = \sigma k^2 (all\ variances\ are\ equal)$
- $H_1 : Atleast\ one\ variance\ is\ different.$

Test Statistic:

$$\chi 2 = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right)}$$

Where

$$N = \sum n_i = totalsamplesize$$

$$s_i^2 = varianceofgroup\ i$$

$$s_p^2 = \frac{\sum(n_i - 1)s_i^2}{N - k} = pooledvariance$$

Degrees of freedom $= k - 1$

Decision Rule:

- $Compare\ \chi^2\ with\ \chi^2\ (k - 1)\ at\ significance\ level\ \alpha.$
- $If \chi^2 calculated > \chi^2 critical \Rightarrow rejectH^0.$

Example:

A researcher wants to check whether 3 different teaching methods produce equal variability in student exam scores.

Method A	72	75	78	74	77
Method B	68	70	65	69	66
Method C	80	85	83	82	84

Step 1: Calculate Variances

Method A

- $Mean\ =\ (72 + 75 + 78 + 74 + 77)/5\ =\ 75.2$
- $Variance\ s_A^2 = 6.7$

Method B

- $Mean\ =\ (68 + 70 + 65 + 69 + 66)/5\ =\ 67.6$
- $Variance\ s_B^2 = 4.3$

Method C

- $Mean\ =\ (80 + 85 + 83 + 82 + 84)/5\ =\ 82.8$
- $Variance\ s_C^2 = 4.7$

So, $s_A^2 = 6.7, s_B^2 = 4.3, s_C^2 = 4.7$

each group size = 5 → $n_1 = n_2 = n_3 = 5$

Total sample size N = 15, groups k = 3

Step 2: Pooled Variance

$$s_p^2 = \frac{(n_1-1)s_A^2 + (n_2-1)s_B^2 + (n_3-1)s_C^2}{N-k}$$

$$s_p^2 = \frac{4(6.7) + 4(4.3) + 4(4.7)}{15-3} = \frac{26.8 + 17.2 + 18.8}{12} = \frac{62.8}{12} = 5.23$$

Step 3: Bartlett's Test Statistic

$$\chi^2 = \frac{(N-k)\ln\left(s_p^2\right) - \sum_{i=1}^{k}(n_i-1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)}$$

- $(N-k) = 12$
- $\ln\left(s_p^2\right) = \ln(5.23) = 1.653$
- $First\ term\ =\ 12 \times 1.653 = 19.84$

  $Now\ log\ terms:$

- $(n_1-1)\ln(s_A^2) = 4\ln(6.7) = 4 \times 1.902 = 7.61$
- $(n_2-1)\ln(s_B^2) = 4\ln(4.3) = 4 \times 1.458 = 5.83$
- $(n_3-1)\ln(s_C^2) = 4\ln(4.7) = 4 \times 1.548 = 6.19$
- $Sum\ =\ 7.61\ +\ 5.83\ +\ 6.19\ =\ 19.63$

Numerator = 19.84 – 19.63 = 0.21

Now correction factor denominator:

$$1 + \frac{1}{3(k-1)}\left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4} - \frac{1}{12}\right) = 1 + \frac{1}{6}(0.75 - 0.083)$$

$$= 1 + \frac{1}{6}(0.667) = 1 + 0.111 = 1.111$$

$So\ test\ statistic:$

$$\chi^2 = \frac{0.21}{1.111} = 0.189$$

Step 4: Decision

- $df = k - 1 = 2$
- $Critical\ value\ at\ \alpha = 0.05: \chi^2_{0.05,2} = 5.99$
- $Our\ value = 0.189 < 5.99$

Conclusion: $Fail\ to\ reject\ H_0.$ So, **the variances of the three teaching methods are equal**.

Advantages:

1) **Powerful under normality**:
   - Bartlett's test is very sensitive in detecting small differences in variances when the data is normally distributed.
2) **Widely used in ANOVA preparation**:
   - Ensures the assumption of homogeneity of variance is met, which is crucial for valid ANOVA results.
3) **Mathematically well-established**:
   - Based on exact distributions under normality, making it theoretically strong.
4) **Useful for multiple groups**:
   - Can compare variances across more than two groups (not limited to pairwise comparisons).

Disadvantages:

1) **Highly sensitive to non-normality**:
   - If the data is not normally distributed, Bartlett's test may give misleading results (false positives/false negatives).
2) **Not robust**:
   - Even slight deviations from normality can cause incorrect conclusions.
3) **Alternatives are better in practice**:
   - **Levene's Test** and **Brown-Forsythe Test** are preferred because they are more robust against non-normal data.

4) **Interpretation depends on sample size**:
   - With large sample sizes, even trivial differences in variances may appear significant.
   - With small samples, it may fail to detect real variance differences.

16.6. LEVENE'S TEST (MODIFIED LEVENE METHOD):

- **Levene's Test** is a statistical test used to check the **homogeneity of variances** (equal variances) across groups, similar to Bartlett's Test.
- It was proposed as a **robust alternative to Bartlett's test**, since Bartlett's is highly sensitive to non-normal data.
- The **Modified Levene's Test** (also called the **Brown–Forsythe version**) uses the **median** instead of the mean to calculate deviations, making it even more robust against skewed data or outliers.
- **Hypotheses**:
  - $H_0 : \sigma^{12} = \sigma^{22} = \cdots = \sigma k^2$(all Population variances are equal)
  - $H_1 : Atleast\ one\ variance\ is\ different.$

Procedure:

1) Compute **absolute deviations** of each observation from its group mean (or median for the Brown–Forsythe version).

$Z_{ij} = |Y_{ij} - \overline{Y}_\iota|$

where Y_{ij} = observation j in group i.

Perform a **one-way ANOVA** on these absolute deviations Z_{ij}.

If the ANOVA is significant → reject H₀, variances are unequal.

- H_0: all variances equal.
- H_1: at least one variance differs.

Problem:

Three groups (each n=5) with these observations:

- **Group A**: 10, 12, 9, 11, 13
- **Group B**: 20, 22, 19, 21, 20
- **Group C**: 30, 40, 28, 35, 32

We want to test

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2$ vs H_a: at least one variance differs.

Use Levene's test (center deviations about the group mean).

Step 1 - Compute Group Means

- Mean $\overline{X_A} = 11.0$
- Mean $\overline{X_B} = 20.4$
- Mean $\overline{X_C} = 33.0$

Step 2 - Compute Absolute Deviations

$Z_{ij} = |Y_{ij} - \overline{Y_\imath}|$

Group A deviations:

|10−11|=1, |12−11|=1, |9−11|=2,|11−11|=0,|13−11|=2|→ **[1, 1, 2, 0, 2]**

Group B deviations:

|20−20.4|=0.4, |22−20.4|=1.6, |19−20.4|=1.4, |21−20.4|=0.6, |20−20.4|=0.4

→ **[0.4, 1.6, 1.4, 0.6, 0.4]**

Group C deviations:

|30−33|=3, |40−33|=7, |28−33|=5, |35−33|=2, |32−33|=1 → **[3, 7, 5, 2, 1]**

Step 3 - Means of the Deviations

- $\overline{Z_A} = 1.20$
- $\overline{Z_B} = 0.88\ (rounded)$

- $\overline{Z_C} = 3.60$

$Overall\ mean\ of\ all\ Z_{ij}: \bar{Z} = 1.8933\ (approx)$

Step 4 - Compute Sums of Squares: One-Way ANOVA on the $\boldsymbol{Z_{ij}}$

Between-group sum of squares (SSB):

$$SSB = \sum_{i=1}^{k} n_i(\bar{Z_\iota} - \bar{Z})^2 = 22.1013 \text{ (approx)}$$

Within-group sum of squares (SSW):

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \bar{Z_\iota}\right)^2 = 27.3280 \text{ (approx)}$$

Degrees of freedom:

$\mathrm{df}_{\text{between}} = k - 1 = 2$

$\mathrm{df}_{\text{within}} = N - k = 15 - 3 = 12$

Mean squares:

$MSB = SSB/(k-1) \ = 11.0507\ (approx)$

$MSW = SSW/(N-k) = 2.2773(approx)$

***F statistic*:**

$$F = \frac{MSB}{MSW} = \frac{11.0507}{2.2773} \approx 4.8525$$

$\boldsymbol{Step5 - Decision(use\ \alpha = 0.05)}$

- $Critical\ F_{0.05,2,12} \approx \mathbf{3.89}.$
- $Observed\ F \approx 4.85 > 3.89\ \rightarrow \boldsymbol{rejectH_0}.$

Conclusion:

There is statistically significant evidence (at the 5% level) that variances are **not equal** across the three groups. In plain words **Group C** has much larger variability than Groups A and B.

Advantages:

1) **Robust to non-normality**:
   - Unlike Bartlett's, Levene's test works well even when data is not normally distributed.
2) **Resistant to outliers** (modified version):
   - By using the **median (Brown–Forsythe modification)**, the test reduces the influence of extreme values.
3) **Applicable for many groups**:
   - Can handle two or more groups easily.
4) **Widely used in ANOVA preparation**:
   - It helps verify the assumption of equal variances before applying parametric tests like ANOVA or t-tests.
5) **Flexibility**:
   - Different versions (based on mean, median, trimmed mean) give options depending on the type of data distribution.

Disadvantages:

1) **Less powerful than Bartlett's under strict normality**:
   - If data is perfectly normal, Bartlett's test is more sensitive in detecting small variance differences.
2) **Sample size issues**:
   - With very small sample sizes, Levene's test may lack power to detect variance differences.
3) **Interpretation with large samples**:
   - Similar to Bartlett's, with very large samples even small, unimportant differences in variances can become statistically significant.
4) **Not fully immune to skewness**:
   - Although better than Bartlett's, extreme skewed distributions can still affect the results.

16.7. APPLICATIONS AND CASE STUDIES:

1) Agriculture

- Crop yield experiments often have heterogeneous variances due to soil quality differences.
- Levene's test is commonly applied.

2) Industrial Experiments

- Machine performance studies: older machines may have higher variability.

3) Medical Research

- Drug response variances may differ between treatment groups.
- Testing equality of variances avoids misleading conclusions.

4) Education Research

- Student performance scores across different teaching methods or schools may have different variability (e.g., private vs. public schools).
- **Levene's test** can be applied before comparing mean scores with ANOVA.

5) Psychology / Social Sciences:

- Reaction times or survey responses often show group differences in variability (e.g., young vs. elderly participants).
- Testing variance equality ensures statistical comparisons (like t-tests or ANOVA) are valid.

16.8. SUMMARY:

Aspect	**Bartlett's Test**	**Levene's Test (Modified Method)**
Purpose	Tests equality of variances (homogeneity of variance) across groups.	Tests equality of variances across groups (robust alternative to Bartlett's).
Hypotheses	H_0: All variances are equal. H_1: At least one variance differs.	H_0: All variances are equal. H_1: At least one variance differs.
Statistic Distribution	Test statistic ~ Chi-square (χ^2) distribution.	Test statistic ~ F-distribution (ANOVA on absolute deviations).
Assumption about Normality	Requires strict normality. Very sensitive to deviations.	Works well under non-normality; robust to skewed data.
Effect of Outliers	Highly sensitive – outliers can distort results.	Modified version (Brown–Forsythe using median) is robust to outliers.

Aspect	**Bartlett's Test**	**Levene's Test (Modified Method)**
Power	More powerful than Levene's if data is perfectly normal.	Slightly less powerful under strict normality, but more reliable in real-world data.
Sample Size Behavior	With small n → less reliable. With large n → even trivial variance differences appear significant.	With small n → reduced power. With large n → may flag small, unimportant differences as significant.
Best Use Case	When data is normal and you want maximum sensitivity.	When data may be non-normal or contain outliers.
Common Applications	Preliminary test for ANOVA when normality is strongly assumed.	Widely used in applied research (psychology, medicine, social sciences) as a standard homogeneity test.

16.9. SELF-ASSESSMENT QUESTIONS:

1) Why is the equal variance assumption necessary in ANOVA?
2) Derive the Bartlett test statistic and explain its limitations.
3) Explain how Levene's test works. Why is it more robust than Bartlett's test?
4) What are the alternatives when homogeneity of variance is violated?
5) Discuss real-life situations where variance heterogeneity occurs.

16.10. SUGGESTED READINGS:

1) Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining - Introduction to Linear Regression Analysis.
2) N.R. Draper and H. Smith - Applied Regression Analysis.
3) Samprit Chatterjee and Ali S. Hadi - Regression Analysis by Example.
4) S.C. Gupta and V.K. Kapoor - Fundamentals of Mathematical Statistics.

Dr. M. Amulya

LESSON-17

ADVANCED MULTIPLE COMPARISON TESTS

17.0. OBJECTIVES:

After completing this unit, you should be able to:

- Understand why multiple comparison tests are required after ANOVA.
- Differentiate between Fisher's LSD, Tukey's HSD, and Duncan's Multiple Range Test.
- Comparison of Methods Fisher's LSD, Tukey's HSD, and Duncan's.
- Apply these tests to identify which group means differ significantly.
- Interpret results in practical research problems.

STRUCTURE:

17.1. INTRODUCTION:

The most useful information from a one-way ANOVA is obtained through examining contrasts. The trick is in picking interesting contrasts to consider. Interesting contrasts are determined by the structure of the treatments or are suggested by the data. The structure of the treatments often suggests a fixed group of contrasts that are of interest. For example, if one of the treatments is a standard treatment or a control, it is of interest to compare all of the other treatments to the standard. With a treatment this leads to a-1 contracts.

One problem is that, with a moderate number of treatment groups, there are many contrasts to look at. When we do tests or confidence intervals, there is a built-in chance for error. The more statistical inferences we perform, the more likely we are to commit an error.

The purpose of the multiple comparison methods examined in this chapter is to control the probability of making a specific type of error. When testing many contrasts, we have many null hypotheses.

This chapter considers multiple comparison methods that control (i.e., limit) the probability of making an error in any of the tests, when all of the null hypotheses are correct. Limiting this probability is referred to as weak control of the experiment wise error rate. It is referred to as weak control because the control only applies under the very stringent assumption that all null hypotheses are correct. Some authors consider a different approach and define strong control of the experiment wise error rate as control of the probability of falsely rejecting any null hypothesis. Thus, strong control limits the probability of false rejections even when some of the null hypotheses are false. Not everybody distinguishes between weak and strong control, so the definition of experiment wise error rate depends on whose work you are reading. One argument against weak control of the experiment wise error rate is that in designed experiments, you choose treatments that you expect to have different effects.

17.2. NEED FOR MULTIPLE COMPARISON TESTS:

Many multiple testing procedures can be adjusted to provide multiple confidence intervals that have a guaranteed simultaneous coverage. Several such methods will be presented.

Besides the treatment structure suggesting contrasts, the other source of interesting contrasts is having the data suggest them. If the data suggest contrast, then the ‘parameter’ in our standard theory for statistical inferences is a function of the data and not a parameter in the usual sense of the word.

When the data suggest the parameter, the standard theory for inferences does not apply. To handle such situations, we can often include the contrasts suggested by the data in a broader class of contrasts and develop a procedure that applies to all contrasts in the class.

In such cases we can ignore the fact that the data suggested particular contrasts of interest because these are still contrasting in the class and the method applies for all contrasts in the class.

17.3. TUKEY’S HONESTLY SIGNIFICANT DIFFERENCE (HSD) TEST:

John Tukey’s honest significant difference method is to reject the equality of a pair of means, say, μ_i and μ_j at the $\alpha = .05$ level, if

$$\frac{|\bar{Y}_{i.} - \bar{Y}_{j.}|}{\sqrt{\frac{MSE}{n}}} > Q\ (0.95, a, df_e)$$

Obviously, this test cannot be rejected for any pair of means unless the test based on the maximum and minimum sample means is also rejected. For an equivalent way of performing the test, reject equality of μ_i and μ_j if

$$|\overline{Y}_i - \overline{Y}_j| > Q\left(1-\frac{\alpha}{2}, a, df_e\right)\sqrt{\frac{MSE}{n}}$$

With a fixed α, the honest significant difference is

$$HSD = Q_{\left(1-\frac{\alpha}{2}, a, df_e\right)}\sqrt{\frac{MSE}{n}}$$

where Q is the studentized range statistic.

For any pair of sample means with an absolute difference greater than the HSD, we conclude that the corresponding population means are significantly different. The HSD is the number that an observed difference must be greater than in order for the population means to have an 'honestly' significant difference. The use of the word 'honest' reflects the view that the LSD method allows 'too many' rejections.

Tukey's method can be extended to provide simultaneous $(1-\alpha)100\%$ confidence intervals for all differences between pairs of means. The interval for the difference $\mu_i - \mu_j$ has end points

$$|\overline{Y}_i - \overline{Y}_j| \pm HSD$$

Where HSD depends on α. For $\alpha = 0.05$, we are 95% confident that the collection of all such intervals simultaneously contains all of the corresponding differences between pairs of population means.

Example:

Using the same fertilizer experiment:

- Fertilizer A: 20, 22, 23
- Fertilizer B: 25, 27, 26
- Fertilizer C: 22, 20, 21

Means

- Fertilizer $A = \frac{20+22+23}{3} = (21.7)$
- Fertilizer $B = \frac{25+27+26}{3} = (26.0)$
- Fertilizer $C = \frac{22+20+21}{3} = (21.0)$

From Comparisons:

- $MSE = 1.0$
- $df_{error} = 6$

- $k = 3,$
- $n = 3$
- $Studentized\, range\ value$:

$$q_{0.05,3,6} \approx 3.67$$

The HSD Formula is given by

$$HSD = Q_{\left(1-\frac{\alpha}{2},a,df_e\right)}\sqrt{\frac{MSE}{n}}$$

$$HSD = 3.67 \times \sqrt{\frac{1.0}{3}} = 2.12$$

Comparisons:

- A vs B = 4.33 → greater than 2.12 → significant
- A vs C = 0.67 → less than 2.12 → not significant
- B vs C = 5.0 → greater than 2.12 → significant

Conclusion:

Fertilizer B is significantly better than A and C, but A and C are not different.

- **Advantages:**
  - Controls overall Type I error.
  - Works best when group sizes are equal.
- **Disadvantage:** Conservative when many comparisons.

17.4. FISHER'S LEAST SIGNIFICANT DIFFERENCE (LSD) METHOD:

The easiest way to adjust for multiple comparisons is to use R.A. Fisher's least significant difference method. To put it as simply as possible, with this method you first look at the analysis of variance. F-test for whether there are differences between the groups. If this test provides no evidence of differences, you quit and go home. If the test is significant at, say, the =05 level, you just ignore the multiple comparison problem and do all other tests in the usual way at the .05 level.

This method is generally considered in appropriate for use with contrasts suggested by the data. While the theoretical basis for excluding contrasts suggested by the data is not clear (at least relative to weak control of the experiment wise error rate), experience indicates that the method rejects far too many individual null hypotheses if this exclusion is not applied. In addition, many people would not apply the method unless the number of comparisons to be made was quite small.

The term least significant difference comes from comparing pairs of means in a balanced ANOVA. There is a number, the least significant difference (LSD), such that the difference between two means must be greater than the LSD for the corresponding treatments to be considered significantly different. Generally, we have a significant difference between μ_i and μ_j if

$$\frac{|\bar{Y}_i - \bar{Y}_j|}{\sqrt{MSE\left[\frac{1}{n}+\frac{1}{n}\right]}} > t\left(1-\frac{\alpha}{2}, df_e\right)$$

Multiplying both sides by the standard error leads to rejection if

$$|\bar{Y}_i - \bar{Y}_j| > t\left(1-\frac{\alpha}{2}, df_e\right)\sqrt{MSE\left[\frac{1}{n}+\frac{1}{n}\right]}$$

The number on the right is defined as the least significant difference,

$$LSD = t_{\left(1-\frac{\alpha}{2}, df_e\right)}\sqrt{\frac{2MSE}{n}}$$

Where

- MSE = Mean Square Error from ANOVA
- n = number of observations per group.

Note that the LSD depends on the choice of but does not depend on which means are being examined. If the absolute difference between two sample means is greater than the LSD the population means are declared significantly different. Recall, however, that these comparisons are never attempted unless the analysis of variance F test is rejected at the level. The reason that a single number exists for comparing all pairs of means is that in a balanced ANOVA the standard error is the same for any comparison between a pair of means.

- If |difference between means| > LSD ⇒ means differ significantly.
- **Advantages:** Simple, powerful.
- **Disadvantage:** Inflates Type I error rate if used without prior ANOVA.

Example:

Suppose we test the effect of 3 fertilizers (A, B, C) on plant growth.

- Fertilizer A: 20, 22, 23
- Fertilizer B: 25, 27, 26
- Fertilizer C: 22, 20, 21

From ANOVA, we get:

Means:

- Fertilizer $A = \frac{20+22+23}{3} = (21.7)$
- Fertilizer $B = \frac{25+27+26}{3} = (26.0)$
- Fertilizer $C = \frac{22+20+21}{3} = (21.0)$

Comparisons:

- $MSE \;=\; 1.0$
- $df_{error} \;= 6$
- $t_{0.05,6} = 2.447$
- $n \;=\; 3$

Fisher's Least Significant Difference Formula:

$$LSD = t_{\left(1-\frac{\alpha}{2}, df_e\right)}\sqrt{\frac{2MSE}{n}}$$

$$LSD = 2.447 \times \sqrt{\frac{2 \times 1.0}{3}} = 2.0$$

Differences:

- A vs B = 4.33 → greater than 2.0 → significant
- A vs C = 0.67 → less than 2.0 → not significant
- B vs C = 5.0 → greater than 2.0 → significant

Conclusion:

Fertilizer B produces significantly more growth than A and C, but A and C do not differ.

17.5. DUNCAN'S MULTIPLE RANGE TEST (DMRT):

Duncan has developed a multiple range procedure similar to that of Newman-Keuls. Newman Keuls uses a series of tabled values

$Q(1-\alpha, a, df_e), Q(1-\alpha, a-1, df_e), \ldots, Q(1-\alpha, 2, df_e)$. Duncan's method simply changes the tabled values. Duncan uses

$Q((1-\alpha)^{a-1}, a, df_e), Q((1-\alpha)^{a-2}, a-1, df_e), \ldots, Q(1-\alpha, 2, df_e)$.

Using Duncan's value $Q[(1-\alpha)^{a-1}, a, df_e]$ to compare the largest and smallest means does not control the experiment wise error rate at α. (It controls it at 1-$(1-\alpha)^{a-1}$). As a result, Duncan suggests performing the analysis of variance Ftest first and proceeding only if the Ftest indicates that there are differences among the means at level.

Duncan's method is more likely to conclude that a pair of means is different than the Newman–Keuls method and less likely to establish a difference than the LSD method. Just as the Newman–Keuls approach can be used to modify the AOM and Dunnett's method, Duncan's idea can also be applied to the AOM and Dunnett's method.

Stepwise Procedure – Compares Ordered Means in Groups.

- Uses a **range statistic** to test differences between ranked means.
- More liberal than Tukey ⇒ greater chance of detecting differences.
- **Advantage:** Higher power.
- **Disadvantage:** Higher risk of Type I error.

Example:

Fertilizer	Plant 1	Plant 2	Plant 3	Mean
A	20	22	23	21.7
B	25	27	26	26.0
C	22	20	21	21.0

From ANOVA, suppose:

$\boldsymbol{Step\ 1\ -\ Order\ means}$

$$C = \frac{22 + 20 + 21}{3} = (21.0)$$

$$A = \frac{20 + 22 + 23}{3} = (21.7)$$

$$B = \frac{25 + 27 + 26}{3} = (26.0)$$

Comparisons

- $MSE\ = \ 1.56$
- $Error\, df\ = \ 6$

$\boldsymbol{Step\ 2\ -\ Calculation}$

$$SE = \sqrt{\frac{1.56}{3}} = 0.72$$

$\boldsymbol{Step\ 3\ -\ Critical\ q - values}$

$Take\ from\ Studentized\ Range\ Table(dependsondf, \alpha = 0.05).$

$For\ r\ = 2, q_2 = 2.95$

$For\ r\ = 3, q_3 = 3.31$

$\boldsymbol{Step\ 4\ -\ Compute\ LSRs}$

- $LSR_2 = 2.95 \times 0.72 = 2.12$
- $LSR_3 = 3.31 \times 0.72 = 2.38$

Step 5 – Compare Differences

- B vs C = 26.0 – 21.0 = 5.0 > 2.38 → Significant
- B vs A = 26.0 – 21.7 = 4.3 > 2.12 → Significant
- A vs C = 21.7 – 21.0 = 0.7 < 2.12 → Not significant

Conclusion: Fertilizer B is significantly better than A and C, but A and C are similar.

17.6. COMPARISON OF METHODS:

Test	Error Control	Power	Suitable When…
Fisher LSD	Weak control	High power	Small number of comparisons, ANOVA significant
Tukey HSD	Strong control	Moderate	All pairwise comparisons, equal sample sizes
Duncan MRT	Moderate	High power	Ordered means, agricultural/biological research

17.7. APPLICATIONS:

- **Agriculture:** Comparing crop yields under different fertilizers.
- **Medicine:** Comparing effects of different drug dosages.
- **Education:** Comparing student performance under different teaching methods.
- **Psychology / Behavioral Science:** Comparing stress levels under different relaxation techniques (e.g., meditation, music therapy, exercise).
- **Manufacturing / Industry:** Comparing the strength of materials produced by different production processes.

17.8. SUMMARY OF MULTIPLE COMPARISON PROCEDURES:

Fisher’s Least Significant Difference (LSD) Test is one of the earliest and simplest multiple comparison procedures. It performs pairwise t-tests between group means but uses the pooled error variance from ANOVA to increase precision. Fisher’s LSD is **powerful** (high ability to detect true differences) because it does not strongly control the familywise Type I error rate when many comparisons are made. As a result, it is generally recommended only when the overall ANOVA F-test is significant, and the number of groups is small. Its primary advantage is sensitivity, but its limitation is **inflation of false positives** in large comparison sets.

Tukey’s Honestly Significant Difference (Tukey’s HSD) Test is one of the most commonly recommended procedures for comparing **all possible pairs of means**. It effectively controls the **familywise error rate**, making it more conservative but also more reliable than Fisher’s LSD, especially when sample sizes are equal. Tukey’s test provides confidence intervals for each mean difference and maintains a strong balance between Type I error control and statistical power. It is the preferred method in many experimental designs where all pairwise comparisons are of interest.

Duncan’s Multiple Range Test (DMRT) is a stepwise, less conservative procedure designed to identify **homogeneous groups of means**. Compared to Tukey’s HSD, Duncan’s test allows more differences to be declared significant because it relaxes Type I error control as the number of steps increases. This gives it greater sensitivity but at the cost of higher false positive rates. Duncan's test is useful when the goal is to maximize detection of group differences, but it is generally not recommended for confirmatory research. Overall, Fisher's LSD is the most liberal, Tukey's offers the best error control, and Duncan's provides greater sensitivity with moderate error protection.

17.9. SELF-ASSESSMENT QUESTIONS:

1) Why do we need multiple comparison tests after ANOVA?
2) Derive the formula for Fisher's LSD test.
3) Differentiate between Tukey's HSD and Duncan's MRT.
4) Discuss advantages and disadvantages of Fisher's LSD.
5) In what situations would you prefer Duncan's MRT over Tukey's HSD?

17.10. SUGGESTED READINGS:

1) Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining - Introduction to Linear Regression Analysis
2) N.R. Draper and H. Smith - Applied Regression Analysis
3) S.C. Gupta and V.K. Kapoor - Fundamentals of Mathematical Statistics

Dr. M. Amulya