U.P. Rajarshi Tandon Open
University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N Linear Models and Design of Experiment

Course Design Committee

Course Preparation Committee

Dr. Shambhavi Mishra **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Editor**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N **LINEAR MODEL & DESIGN OF EXPERIMENT**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Blocks & Units Introduction

The present SLM on ***Linear Model & Design of Experiment*** consists of nine units with three blocks.

The ***Block - 1 – Linear Estimation and Analysis of Variance,*** is the first block, which is divided into three units.

The ***Unit - 1 – Linear Model and BLUE,*** is the first unit of present self-learning material, which describes Linear Estimation- estimable functions, estimations and error space, Best linear unbiased estimate (BLUE), Markov theorem distribution of quadratic form, Estimable linear hypotheses generalized F and T tests.

In ***Unit – 2 – Analysis of Variance- I,*** the main emphasis on the Analysis of Variance : one-way and two-way classification with equal number of observation per cell and analysis with missing observations.

In ***Unit – 3 – Analysis of Variance- II,*** we have focussed mainly on Analysis of Variance: one-way and two-way classification with unequal number of observations per cell, analysis with missing observations, Tukey's test general two-way classification, Analyses of covariance.

The ***Block - 2 –Design of Experiment*** is the second block with three units.

In ***Unit – 4 – Basic Designs,*** is being introduced the Terminology and basic Principles of Design, CRD, RBD and LSD, analysis with missing observations.

In ***Unit – 5 – Factorial Experiments*** is discussed with 2^3 , 2^n , 3^2 and 3^3 factorial experiments with its analysis.

In ***Unit – 6 – Confounding*** has been introduced, Orthogonality, Complete and Partial confounding, construction of confounded factorial experiments.

The ***Block - 3 – Advance Theory of Design of Experiment*** has three units.

Unit – 7 – BIBD and PBIBD dealt with Balanced Incomplete Block Design (BIBD), Partially Balanced Incomplete Block Design (PBIBD), construction of BIBD and PBIBD, association schemes and construction, resolvable and affine resolvable design.

Unit – 8 – Split and Strip Plot Design, comprises the Intra block and inter block analysis, Split Plot Design, Strip Plot Design.

In ***Unit – 9 – Other Advance Design***, we have discussed the Dual and linked block design, Lattice Designs, Cross-over designs, optimal designs- optimal criteria, robust parameter design, response surface design – orthogonality, rotatability and blocking, weighing designs, mixture experiments

At the end of every block/unit the summary, self-assessment questions and further readings are given.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N Linear Models and Design of Experiment

Block: 1 Linear Estimation and Analysis of Variance

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Shambhavi Mishra **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Editor**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N **LINEAR MODEL & DESIGN OF EXPERIMENT**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Block & Units Introduction

The ***Block - 1 – Linear Estimation and Analysis of Variance,*** is the first block of said SLM, which is divided into three units.

The ***Unit - 1 – Linear Model and BLUE,*** is the first unit of present self-learning material, which describes Linear Estimation- estimable functions, estimations and error space, Best Linear Unbiased Estimate (BLUE), Markov theorem distribution of quadratic form, Estimable linear hypotheses generalized F and T tests.

In ***Unit – 2 – Analysis of Variance- I,*** the main emphasis on the Analysis of Variance: one-way and two-way classification with equal number of observations per cell and analysis with missing observations.

In ***Unit – 3 – Analysis of Variance- II,*** we have focussed mainly on Analysis of Variance: one-way and two-way classification with unequal number of observations per cell, analysis with missing observations, Tukey’s test general two-way classification, Analyses of covariance.

At the end of every unit the summary, self-assessment questions and further readings are given.

UNIT-1 LINEAR MODEL AND BLUE

Structure

1.1 Introduction

Regression analysis is designed for situations where a researcher thinks that a variable is related to one or more other measurements made, usually on the same object. A purpose of the analysis is to use data (observed values of the variables) to estimate the form of this relationship. An example would be to use information on income and the number of years of formal schooling (beyond the sixth grade) to estimate the extent to which a person's annual income is related to his/her years of schooling. One possibility is that for a person with zero years beyond sixth grade, a researcher would anticipate an income of a rupee. For each year of schooling beyond sixth grade, a person has the researcher would expect that his/her income would be larger by b rupees. Thus, for a person with x years of schooling beyond sixth grade, the researcher would expect an annual income of a + bx rupees. When we say that the researcher would expect an annual income of a + bx rupees, we refer to the average income of all people that have had x years of school beyond sixth grade. If y denotes income and x denotes years of schooling beyond sixth grade, we write E(y) for expected income. This leads to the relationship

$$E(y) = a + bx \tag{1}$$

The attempted description of how we think one variable is related to another variable is an example of what is called *model building*. The model here that a person's income is expected to be a + bx where x is his/her number of years of schooling beyond sixth grade is a linear model because we envisage E(y) as being a linear combination of the unknowns a and b. These unknowns are called *parameters*.

1.2 Objectives

After going through this unit, you should be able to:

- Understand the basic concepts of linear estimations about the model building, various properties, estimable functions etc.,
- Obtain the Best Linear Unbiased Estimator (BLUE) for the full-rank model,
- Use Markov theorem distribution of quadratic form,
- Test the general hypotheses of linear estimation using generalized F and t tests.

1.3 Linear Model

A model is termed as linear if it is linear in terms of parameters i.e., if the partial derivative of y with respect to each of the parameter $b_1, b_2 \ldots . b_k$ are independent of the parameters. Linearity of the model is not described by the linearity of the explanatory variable:

Example:

1. $y = b_1 x^2 + b_2 \sqrt{x_2} + b_3 \sqrt{(\log x_3)} + e$ is a linear model because $\frac{\delta y}{\delta b_i} (i = 1,2,3)$ are independent of b_1, b_2, b_3
2. $y = b_1^2 + b_2 x_2 + b x_3 + e$ is a nonlinear model because $\frac{\delta y}{\delta b_1} = 2 b_1 x_1$

If in general f is chosen as:

$$f(x_1, x_2 \ldots \ldots . . x_k, b_1, \ldots \ldots b_k) = b_1 x_1 + b_2 x_2 + \cdots \ldots \ldots \ldots . + b_k x_k$$

to describe a linear model. The aim of statistical linear modelling is to determine $b_1, b_2 \ldots \ldots b_k$ given by the observation on y and $x_1, x_2 \ldots \ldots . . x_k$.

Consider a simple linear regression model

$$y_i = a + bx_i + e_i \quad (2)$$

Where y_i is dependent or study variable and x_i is termed as the independent or explanatory variable. The term $a \, and \, b$ are the parameter or regression coefficient of the model. The parameter a is termed as intercept term and b is termed as the slope parameter. e_i is the error term which is identically and independently distributed random variable with mean zero and constant variance σ^2.

The term e_i represents the extent to which an observed yi differs from its expected value, i.e., $e_i = y_i - (a + bx_i)$. The characteristics of $e_i; s$ are:

a. The expected value of e_i are zero, i.e., $E(e_i) = 0$
b. The variance of e_i is σ^2 for all i, i.e., $V(e_i) = E[e_i - E(e_i)]^2 = \sigma^2$
c. The covariance between any pairs of e_i is zero, i.e., $Cov(e_i e_j) = E[\{(e_i - E(e_i)\}\{(e_j - E(e_j)\}] = 0, for\ all\ i \neq j$

1.3.1 Least Square Estimation

There are several well-recognized methods that can be used for estimating a and b. A frequently used method is known as least squares. Least-squares estimation involves minimizing the sum of the squares of the deviations of the observed yi's from their expected values.

Suppose a sample of *n* sets of paired observations (x_i, y_i) $(i = 1,2 \ldots . n)$ is available. These observations are assumed to satisfy the simple linear regression model, and so we can write

$$y_i = a + bx_i + e_i \ (i = 1,2 \ldots .. n)$$

The principle of least squares estimates the parameters a and b by minimizing the error sum of square.

a) When error is vertical difference, then method is known as Direct Regression
b) When error is horizontal difference, then method is known as Reverse Regression
c) When error is perpendicular distance, then method is known as Orthogonal Regression.

Generally, the direct regression approach estimates are referred as the least square estimates or ordinary least squares estimates.

Direct Regression Method

Assuming that a set of n paired observations on (x_i, y_i) $(i = 1,2 \dots . n)$are available which satisfy the linear regression model. So, we can write the model for each observation as:

$$y_i = a + bx_i + e_i \; (i = 1,2 \dots . . n)$$

Using the direct regression approach minimizing the error sum of the square we get:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad (3)$$

Now taking the partial derivatives of (3) with respect to a is:

$$\delta SSE/\delta a = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$

And the partial derivative of SSE with respect to β_1 is

$$\delta SSE/\delta b = -2 \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i$$

The solution of a and b are obtained by setting $\delta SSE/\delta a = 0$ and $\delta SSE/\delta b = 0$

The solution of these two equations is called the ***direct regression estimators***, or usually called as the ***Ordinary Least Square (OLS) Estimators*** of a and b.

This gives the ordinary least square estimates $\hat{a}$ of a and $\hat{b}$ of b as:

$$\hat{a} = \bar{y} - b\bar{x}$$

$$\hat{b} = \frac{S_{(xy)}}{S_{(xx)}}$$

Where,

$S_{(xy)} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, $S_{(xy)} = \sum_{i=1}^{n} (x_i - \bar{x})^2$,

$\bar{x} = 1/\sum_{i=1}^{n} x_i$, $\bar{y} = 1/\sum_{i=1}^{n} y_i$.

1.3.1.1 Properties of Ordinary Least Square Estimators

1. Unbiased Property

$$\begin{aligned} E(a) &= E(\bar{y} - b\bar{x}) \\ &= E\left(a + b\bar{x} - \hat{b}\bar{x}\right) \\ &= a + b\bar{x} - \bar{x}E\left(\hat{b}\right) \\ &= a \end{aligned}$$

$E(\hat{b}) = b$

â and $\hat{b}$ are unbiased estimators of a, b respectively.

2. Variance

$var(\hat{b}) = \frac{\sigma^2}{S_{(xx)}}$

$var(\hat{a}) = var(\bar{y} - \hat{b}\bar{x})$

$= var(\overline{y_n}) + \overline{x^2}v(\hat{b}) - 2\bar{x}cov(\hat{b}, \bar{y})$

$= \frac{\sigma^2}{n} + \frac{\overline{x^2}}{S_{(xx)}} * \sigma^2 + 0$

$= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x^2}}{S_{(xx)}} \right)$

$\mathrm{var}(a) = \frac{\sigma^2}{S_{(xx)}} * \frac{1}{n} \sum x_i^2$

3. Covariance

$cov(\hat{a}, \hat{b}) = cov(\bar{y} - b) - \bar{x}var(b)$

$cov(\hat{a}, \hat{b}) = \frac{-\bar{x}\sigma^2}{S_{(xx)}}$

4. Residual Sum of Square (RSS)

$SS_{(res)} = \sum_{i=1}^{n}(\overline{y_1} - \hat{y_l})^2$

$= \sum_{i=1}^{n}(y_i - a - bx_i)^2$

$= \sum_{i=1}^{n}(y_i - \bar{y} + b\bar{x} - bx_i)^2$

$= \sum_{i=1}^{n}[(y_i - \bar{y}) - b(x_i - \bar{x})^2]$

$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + b^2\sum_{i=1}^{n}[(x_i - \bar{x})^2 - 2b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$= S_{(yy)} - b^2 S_{(xx)} - 2b^2 S_{(xx)}$

$= S_{(yy)} - b^2 S_{(xx)}$

$= S_{(yy)} - \left(\frac{S_{(xy)}}{S_{(xx)}}\right)^2 S_{(xx)}$

$= S_{(yy)} - \frac{S_{(xy)}^2}{S_{(xx)}}$

$= S_{(yy)} - bS_{(xy)}$

Where $S_{(yy)} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Estimation of σ^2

The estimator of σ^2 is obtained from the residual sum of squares as follows. Assuming that y_i is normally distributed, it follows that $SS_{(res)}$ has a χ^2 distribution with $(n-2)$ degree of freedom, so:

$$\frac{SS_{(res)}}{\sigma^2} \sim \chi^2(n-2)$$

Thus, using the result about the expectation of a chi-square random variable, we have:

$$E(SS_{(res)} = (n-2)\sigma^2$$

Thus, an unbiased estimator of σ^2 is:

$$s^2 = \frac{SS_{(res)}}{n-2}$$

1.3.2 General Case of *k* Variables

The linear model equation is represented as

$$y = Xb + e \text{ with } E(y) = Xb$$

thus, for k variables

$$X = \begin{bmatrix} 1 & x_{11} & \cdots\cdots & x_{1k} \\ 1 & x_{21} & \cdots\cdots & x_{2k} \\ \vdots & \vdots & \cdots\cdots & \vdots \\ \vdots & \vdots & \cdots\cdots & \vdots \\ 1 & x_{N1} & \cdots\cdots & x_{Nk} \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ \vdots \\ b_k \end{bmatrix}$$

and y and e defined as above are unchanged

as we know that $E(e) = 0$ and $var(e) = E[e - E(e)][e - E(e)]' = E(ee') = \sigma^2 I_N$

by the method of least square estimation

$e'e = [y - E(y)]'[y - E(y)] = (y - Xb)'(y - Xb)$

$= y'y - 2b'X'y + b'X'Xb$

To obtain the estimator $\hat{b}$, that value of b that minimizes e'e ,we must differentiate e'e with respect to the elements of b and setting the result equal to zero.

$\frac{\delta(e\prime e)}{\delta b} = -2X'y + X'Xb = 0$

$-2X'y + X'Xb = 0$

$X'Xb = X'y$ (4)

The equation (4) is known as the normal equations. Provided $(X'X)^{-1}$ exists, they have a unique solution for $\hat{b}$

$$\hat{b} = (X'X)^{-1}X'y \quad (5)$$

When $X'X$ is nonsingular (of full rank) the unique solution of (4) can be written as (5).

When $X'X$ is singular, the solution will take the form:

$$\hat{\mathrm{b}} = \mathrm{G}\mathrm{X}'\mathrm{y} \quad (6)$$

Where G is a generalized inverse of $X'X$. This solution is not unique because generalized inverses are not unique.

By the nature of X, $X'X$ is square of order $k + 1$ with elements that are sums of square and products and $X'y$ is the vector of sums of products of the observed $x's\ and\ y's$. As a result, we have:

$$X'y = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i1}y_i \\ \vdots \\ \vdots \\ \sum_{i=1}^{N} x_{ik}y_i \end{bmatrix}$$

$$\mathrm{X'X} = \begin{bmatrix} \mathrm{N} & \mathrm{x.1} & \mathrm{x.2} & \cdots\cdots\cdots & \mathrm{x.k} \\ \mathrm{x.1} & \sum_{i=1}^{N} x_{i1}^2 & \sum_{i=1}^{N} x_{i1}x_{i2} & \cdots\cdots\cdots & \sum_{i=1}^{N} x_{i1}x_{ik} \\ \mathrm{x.2} & \sum_{i=1}^{N} x_{i1}x_{i2} & \sum_{i=1}^{N} x_{i2}^2 & \cdots\cdots\cdots & \sum_{i=1}^{N} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{x.k} & \sum_{i=1}^{N} x_{i1}x_{ik} & \sum_{i=1}^{N} x_{i2}x_{ik} & \cdots\cdots\cdots & \sum_{i=1}^{N} x_{ik}^2 \end{bmatrix}$$

Method of Estimation

In obtaining the least-square estimator, we shall assume a model of the form y = Xb + e where X has full column rank, E(y) = Xb, and E(e) = 0. To obtain an alternative to the least-square estimator, we shall assume that b is a random variable with a known mean and covariance matrix.

1. **Ordinary Least Square Estimation**

This involve choosing $\hat{b}$ as the value of b which minimizes the sum of squares of observations from their expected values,

$$\sum_{i=1}^{N}[y_i - E(y)]^2 = (y - Xb)'(y - Xb)$$

The resulting estimator is as we have seen:

$$\hat{b} = (X'X)^{-1}X'y$$

2. **Generalized Least Squares**

This is also called weighted least squares. Assume that the variance covariance matrix of e is var(e) =V. Now minimize $= (y - Xb)'V^{-1}(y - Xb)$ with respect to b. The resulting estimator is:

$$\hat{b} = (X'V^{-1}X)^{-1}\, X'V^{-1}y$$

When it is assumed that the components of $var(\boldsymbol{e})$ are equal and uncorrelated, that is**,** $\boldsymbol{V} = \sigma^2 I$, the generalized or weighted least estimators reduced to the ordinary least-square estimators.

1.3.3 Best Linear Unbiased Estimator (BLUE)

When the least square estimator is the linear unbiased estimator of the parameters of a regression that has minimum variance, then it is called as Best Linear Unbiased Estimator (BLUE).

For any row vector t' with the same number of columns as there are rows of b ,the scalar $t'b$ is a linear function of the elements of the vector of parameters b.

The three characteristics of the estimator under study are linearity, unbiasedness, and being the best estimator (the one with the smallest variance).

i) **Linearity:** The estimator is to be a linear function of the observations y. Let this estimator be $\lambda' y$ where λ' is a row vector of order N. We shall show that λ is uniquely determined by the other two characteristics of the estimator.

ii) **Unbiasedness:** The estimator $\lambda' y$ is to be unbiased for $t'b$. Therefore, we must have that $E(\lambda' y) = t'b$. However, $E(\lambda' y) = \lambda' Xb$ so that $\lambda' Xb = t'b$. Since this must be true for all b, we have that:

$$\lambda' X = t'. \qquad (7)$$

iii) **A best estimator:** Here, "*best*" means that in the class of linear, unbiased estimators of $t'b$, the best is the one that has minimum variance. This is the criterion for deriving λ'.

1.4 Gauss – Markov Theorem

Assume that for the linear model:

$$y = Xb + e,$$

$var(e) = V$. Then the best linear unbiased estimator of $t'b$ is:

$$t'\hat{b} = t'(X'V^{-1}X)X'V^{-1}y$$

Proof:

Since $var(e) - V, var(y) = V$.

Then $\mathrm{var}(\lambda' y) = \lambda' V\lambda$

We must minimize this quantity with respect to the constraint $(\lambda' X = t')$ in (7). To do this, we use the method of Lagrange multipliers. Using 2θ as a vector of Lagrange multipliers, we therefore minimize:

$$w = \lambda' V\lambda - 2\theta'(X'\lambda - t)$$

with respect to the elements of $\lambda'$and θ'. We differentiate w with respect to θ., set it equal to zero and get (7). Differentiation of w with respect to λ gives:

$V\lambda = X\theta$ or $\lambda = V^{-1}\theta$

Since V^{-1} exists. Substitution in (7) gives $t' = \lambda' X = \theta' X' V^{-1}X$ and so $\theta' = t'(X' V^{-1}X)^{-1}$

Hence,

$$\lambda' = \theta' X' V^{-1} = t'(X'V^{-1}X)^{-1}X' V^{-1} \qquad (8)$$

The BLUE of $t'b$ is:

$$t'b = t'(X'V^{-1}X)^{-1}X' V^{-1}y$$

We have shown that the BLUE is weighted or generalized least square estimators. Its variance is $var\left(t'\hat{b}\right) = t'(X'V^{-1}X)^{-1}t$

Since (8) is the sole solution to the problem of minimizing $var(\lambda' y) = \lambda' V\lambda$ Subject to constraint (7), the BLUE $\lambda' y$ of $t'b$ is the unique estimator of $t'b$ having the properties of linearity, unbiasedness, and "*bestness*"—minimum variance of all linear unbiased estimators. thus, the BLUE of $t'b$ is unique $\lambda' y\ for\ \lambda'$ as given in (8).

1.5 Estimable Functions

A linear function of the parameters is defined as *estimable* if it is identically equal to some linear function of the expected value of the vector of observations. This means that $q'b$ is estimable if $q'b = t'E(y)$ for some vector t'. In other words, if a vector t' exists such that, $t'E(y) = q'b$ then $q'b$ is said to be estimable. Note that in no way is there any sense of uniqueness about t'. It simply has to exist.

1.5.1 Some Properties of Estimable Functions

i) The Expected Value of Any Observation is Estimable

The definition of an estimable function is that $q'b$ is estimable if $q'b = t'E(y)$ for some vector t'. Consider a t' which has one element unity and the others zero. Then, $t'E(y)$ will be estimable. It is an element of $E(y)$ the expected value of an observation. Hence, the expected value of any observation is estimable.

ii) Linear Combinations of Estimable Functions are Estimable

Every estimable function is a linear combination of the elements of E(y).This is also true about a linear combination of estimable functions. Thus, a linear combination of estimable functions is also estimable. More formally, if $q_1'b$and $q_2'b$and are estimable, there exists a $t_1'$and at_2' such that $q_1'b = t_1'E(y)$ and $q_2'b = t_2'E(y)$ Hence, a linear combination $c_1q_1'b + c_2q_2'b = (c_1t_1' + c_2t_2')E(y)$and so it is estimable.

iii) The Forms of an Estimable Function

If $q'b$is estimable using its definition, we have that for some vector t'

$$q'b = t'E(y) = t'E(Xb) = t'Xb \qquad (1)$$

Since estimability is a concept that does not depend on the value of b, the result in equation (1) must be true for all b.

Therefore, $q' = t'X$ for some vector t' (2)

iv) Invariance to the Solution b°

When $q'b$ is estimable, $q'b^o$ is invariant to whatever solution of the normal equations $X'Xb^o = X'y$ is used for b_0. If $q'b$ is estimable, then $q'b^o$has the same value for all solutions b_0 to the normal equations.

Example: Consider the model $y_1 = \mathrm{b}_1 + \mathrm{b}_2 + e_1$, $y_2 = \mathrm{b}_1 + \mathrm{b}_3 + e_2$ and $y_3 = \mathrm{b}_1 + \mathrm{b}_2 + e_3$. Show that $q_1b_1 + q_2b_2 + q_3b_3$ is estimable if and only if $b_1 = b_2 + b_3$

Solution: Consider a linear function $a_1b_1 + a_2b_2 + a_3b_3$ is such that its expectation is $q_1b_1 + q_2b_2 + q_3b_3$ identically. Then

$E(a_1b_1 + a_2b_2 + a_3b_3) = a_1(b_1 + b_2) + a_2(b_1 + b_3) + a_3(b_1 + b_2)$

$$= (a_1 + a_2 + a_3)b_1 + (a_1 + a_3)b_2 + a_2b_3$$

And if this $= \sum_{i=1}^{3} q_ib_i$,we have:

$q_1 = a_1 + a_2 + a_3, q_2 = a_1 + a_3, q_3 = a_2$

And therefore, $a_1 = a_2 + a_3$

Conversely if $a_1 = a_2 + a_3$

$q_1b_1 + q_2b_2 + q_3b_3 = (q_1 + q_3)b_1 + q_2b_2 + q_3b_3$

$$= q_2(b_1 + b_2) + b_3(b_1 + b_2)$$

$$= q_2E(y_1) + q_3E(y_2)$$

And hence there exists a function $q_2y_1 + q_3y_2$ whose expectation is $\sum_{i=1}^{3} q_ib_i \ or \ q_ib_i$ is estimable.

1.5.2 Estimation Space and Error Space

The space generated by the column vector of X is called the estimation space of X and is denoted by $V(x_1, x_2 \ldots\ldots x_p) = \mathrm{V_p}$, obviously $\mathrm{V_p} \subset R_n(because\ , p \leq k)$

$\tau = (\tau_1, \ldots\ldots\ldots \tau_2)'$belongs to $\mathrm{V_p}$,if it is linear combination of $x_i's$ i.e., if

$\tau = \sum_{j=1}^{p} \theta_j x_k$ for some real $\theta_1, \theta_2 \ldots \ldots \ldots \ldots \theta_p$

Thus, $\mathrm{V_p} = \theta_1 x_1 + \cdots \ldots \ldots \ldots .. + \theta_p x_p | \theta_1 \ldots \ldots \ldots .. \theta_p \subset R^1$

Clearly, $\eta \in \mathrm{V_p}$ .Rank of estimation space is the number of independent vectors among $x_1, x_2 \ldots \ldots x_p$

The vector space E that is orthogonal to the vector space V_p is called the error space. Thus if γ is any vector in E, $\gamma' x_j = 0$ for $\forall j = 1 \ldots .. p$

Example: $E(y) = b_0 1 + b_1\, x$ where x $= (x_1, x_2 \ldots . x_n)'$. Here $X = (1\ x)$.
Estimation space $V_2 = \tau = \theta_1 1 + \theta_2 x | \theta_1, \theta_2 \in R_1 \subset R^n$ and contains the point $\eta = b_0 1 + b_1\, x$. The error space is space which is orthogonal to V_2i.e., $E = \{a : a'1 = 0, a'x = 0\}$.

1.6 Gauss – Markov Theorem for Quadratic Form

The best linear unbiased estimator of the estimable function $q'b$ is $q'b^o$; that is,

$$\widehat{q'b} = q'b^o \quad (3)$$

where by the "hat" notation we mean "*BLUE of*".

Proof:

Form the property of linearity, unbiasedness, and "bestness" (having minimum variance).

Since $q'b^o$ is a linear function of the observation, because $q'b^o = q'GX'y$

And $q'b^o$ is an unbiased estimator of $q'b$ because

$$E(q'b^o) = q'E(b^0) = q'GX'Xb = t'XGX'Xb = t'Xb = q'b \quad (4)$$

In establishing equation (4), we invoke:

$$q'b^0 = d'U'GX'y = d'U'GU\Lambda^{1/2}Sy = d'\Lambda^{1/2}\Lambda^{-1/2}\mathrm{Sy} = \mathrm{d}'\mathrm{Sy}$$

and, when G is a generalized inverse of $X'X$ then

$$X = XGX'X \text{ which also implies } X'XGX \quad (5)$$

Alternatively, when the singular value decomposition of $X = S'\Lambda^{1/2}U'$ then

$$X'X = U\Lambda\mathrm{U}' \text{ and } (\mathrm{X}'\mathrm{X})^{+} = \mathrm{U}\Lambda^{-1}U'$$

and for any generalized inverse of $X'X, U'(\mathrm{X}'\mathrm{X})^{-}U = \Lambda^{-1}$and therefore:

$$(\mathrm{X}'\mathrm{X})^{+} = UU'(\mathrm{X}'\mathrm{X})^{-}UU'$$

Thus, $E(q'b^o) = q'E(b^o) = q'GX'Xb = d'U'GU\Lambda U'b = d'\Lambda^{-1}\Lambda\mathrm{U}'\mathrm{b} = \mathrm{d}'\mathrm{U}'\mathrm{b} = \mathrm{q}'\mathrm{b}$

To demonstrate that $q'b^0$ is a best estimator, we need its variance. We then show that the variance of any other linear unbiased estimator of $q'b$ is larger. We have that:

$$v(q'b^0) = q'GX'XG'q\sigma^2$$

$$= q'GX'XG'Xt\sigma^2$$

$$= q'GX't\sigma^2$$

$$= q'Gq\sigma^2 \quad (6)$$

Using the equation (5), we now show that $q'b^0$ has the minimum variance among all the linear unbiased estimators $q'b$ and hence is the best. Suppose that $K'y$ is unbiased,$E(K'y) = q'b$ so $k'X = q'$. Therefore:

$$cov(q'b^0, k'y) = cov(q'GX'y, k'y) = q'GX'k\sigma^2 = q'Gq\sigma^2$$

Consequently,

$$v(q'b^0 - k'y) = v(q'b^0) + v(k'y) - 2cov(q'b^0, k'y)$$

$$= v(k'y) - q'Gq\sigma^2$$

$$= v(k'y) - v(q'b^0) > 0 \qquad (7)$$

Since $v(q'b^0 - k'y)$ is positive, from equation (7), $v(k'y)$ exceeds $v(q'b^0)$. Thus $q'b^0$ has a smaller variance that any other linear unbiased estimator of $q'b$ and hence is the best.

If $q'b$ is an estimable function, its BLUE is $q'b^0$ with variance $q'Gq\sigma^2$. This is so for any solution b^0 to the normal equations using any generalized inverse G . Both the estimator and its variance are invariant to the choice of $G \ and \ b^0$.

Similarly, the covariance between the BLUEs of two estimable functions

$$cov(q_1'b^0, q_2'b^0) = q_1'Gq_2\sigma^2$$

Hence, if $Q'b^0$ represent the BLUEs of several estimable functions, the variance covariance matrix of these BLUE's is $var(Q'b^0) = Q'GQ\sigma^2$

1.7 General Linear Hypothesis

In testing of linear hypothesis, four hypotheses of particular interest are:

(i) H: b = 0, the hypothesis that all of the elements of b are zero;

(ii) H: $b = b_0$, the hypothesis that $b_i = b_{i0}$ for i = 1, 2,…, k, that is, that each b_i is equal to some specified value b_{i0};

(iii) H: λ' b = m, that some linear combination of the elements of b equals a specified constant;

(iv) H: $b_q = 0$, that some of b_i's, q of them where q < k is zero.

To conduct these hypotheses, we need certain assumptions:

1) When $x{\sim}N(\mu, V)$, the quadratic forms $x'Ax$ and $x'Bx$ are distributed independently if and only if $AVB = 0$ (or equivalently $BVA = 0$)

2) The matrix $A'A$ is positive definite when A has full-row rank is positive-semi-definite otherwise

All of the linear hypothesis above and others are special cases of a general procedure even though the calculation of the F-statistics may appear to differ from one hypothesis to another.

In general hypothesis we consider is:

$$H: K'b = m$$

Where, b ,is the $(k+1)$ order vector of parameters of the model, K' is any matrix of s rows and k+1 columns and m is a vector of order s of specified constants. K' must be full row rank *i.e.*, $r(K') = s$ means that the linear functions of b must be linearly independent. The hypothesis being tested must be made up of linearly independent functions of b and must contain no functions that are linear functions of others therein.

We now develop the F-statistic to test the hypothesis H: K′ b = m.

We know that:

$$y \sim N(Xb, \sigma^2 I), \hat{b} = (X'X)^{-1}X'y \text{ and } \hat{b} \sim N[(X'X)^{-1}\sigma^2]$$

Therefore,

$$K'\hat{b} - m \sim N[K'b - m, K'(X'X)^{-1}K\sigma^2]$$

By the Assumption 1), the quadratic form:

$$Q = (K'\hat{b} - m)'[K'(X'X)^{-1}K]^{-1}(K'\hat{b} - m)$$

In $= \left(K'^{\hat{b}} - m\right)$ with matrix $[K'^{(X'X)^{-1}}K]^{-1}$ has a non-central χ^2-distribution. We have that:

$$\frac{Q}{\sigma^2} \sim \chi^{2\prime}\{s, \frac{(K'\hat{b}-m)'[K'^{(X'X)^{-1}}K]^{-1}(K'\hat{b}-m)}{2\sigma^2}\} \quad (1)$$

We now show the independence of Q and SSE using Assumption 1), we first express Q and SSE as quadratic forms of the same normally distributed random variable. We note that the inverse of $K'(X'X)^{-1}K$ exists because K' has full row rank and $X'X$ is symmetric.

Now, in (1), we replace $\hat{b}$ with $(X'X)^{-1}X'y$. Then (1) for Q becomes:

$$Q = (K'(X'X)^{-1}X'y - m)'[K'(X'X)^{-1}K]^{-1}(K'(X'X)^{-1}X'y - m)$$

The matrix K' has full-column rank. Assumption 2), $K'K$ is positive definite. Thus $K'K^{-1}$exists.

Therefore,

$$(K'(X'X)^{-1}X'y - m) = K'(X'X)^{-1}X'[y - XK(K'K)^{-1}m]$$

As a result, Q may be written:

$$Q = [y - XK(K'K)^{-1}m]'\, X'(X'X)^{-1}K[K'(X'X)^{-1}K]^{-1}K'(X'X)^{-1}X'[y - XK(K'K)^{-1}m]$$

The next step is to get the quadratic form for SSE into a similar form as Q:

$$SSE = y'[I - X(X'X)^{-1}X']y$$

Since, $X'[I - X(X'X)^{-1}X'] = 0$ and $[I - X(X'X)^{-1}X']X = 0$,we may write

$SSE = [y - XK(K'K)^{-1}m]'[I - X(X'X)^{-1}X'][y - XK(K'K)^{-1}m]$

We have expressed both Q and SSE as quadratic forms in the normally distributed vector $y - XK(K'K)^{-1}m$. Also, the matrices for Q and SSE are both idempotent, so we again verify that they have χ^2-distribution. More importantly, the product of the matrices for Q and SSE are null. We have that:

$$[I - X(X'X)^{-1}X]'\, X'(X'X)^{-1}K[K'(X'X)^{-1}K]^{-1}K'(X'X)^{-1}X'$$

Therefore, by assumption 1) Q and SSE are distributed independently. This gives us the F-distribution needed to test the hypothesis $H: K'b = m$. We have that:

$$F(H) = \frac{Q/_{s}}{SSE/_{[N-r(X)]}} = \frac{Q}{s\widehat{\sigma^2}}$$

$$\sim F'(s, N - r(X), \frac{(K'\mathrm{b}-m)'[K'\left(X'X\right)^{-1}K]^{-1}(K'\mathrm{b}-m)}{2\sigma^2} \qquad (2$$

Under the null hypothesis $H: K'b = mF(H) \sim F_{(s,N-r(X))}$, Hence, $F(H)$ provides a test of the null hypothesis is $H: K'b = m$ and the F-statistics for testing this hypothesis is:

$$F(H) = \frac{Q}{s\widehat{\sigma^2}} = \frac{(K'\mathrm{b}-m)'[K'\left(X'X\right)^{-1}K]^{-1}(K'\mathrm{b}-m)}{s\widehat{\sigma^2}} \qquad (3)$$

With s, and N-r degree of freedom.

The generality of this result merits emphasis. It applies for any linear hypothesis $K'b = m$,The only limitation is that K' has full-row rank. Other than this $F(H)$can be used to test any linear hypothesis whatever. No matter what the hypothesis is, it only has to be written in the form $K'b = m$,. Then, $F(H)$of equation (3) provides the test. Having once solved the normal equations for the model $y = Xb + e$ and so obtained$(X'X)^{-1}$, $\hat{b} = (X'X)^{-1}X'y$ and $\widehat{\sigma^2}$ the testing of $H: K'b = m$ can be achieved by immediate application of $F(H)$.

1.7.1 Estimation under the Null Hypothesis

By the least square method $\widehat{b_c}$ is derived so as to minimize $\left(y - X\widehat{b_c}\right)'\left(y - X\widehat{b_c}\right)$ subject to the constraint $K'b = m$

With $2\theta'$ as a vector of Lagrange multipliers, we minimize:

$$L = \left(y - X\widehat{b_c}\right)'\left(y - X\widehat{b_c}\right) + 2\theta'\left(K\widehat{b_c} - m\right)$$

With respect to the elements of $\widehat{b_c}$ and θ'.Differentiation with respect to these elements leads to the equations:

$$X'X\widehat{b_c} + K\theta = X'y \qquad (4)$$

$$K'\widehat{b_c} = m$$

From these two equations:

$$\widehat{b_c} = (X'X)^{-1}(X'y - K\theta) = \hat{b} - (X'X)^{-1}K\theta \qquad (5)$$

And $$K'\widehat{b_c} = K'\hat{b} - K'(X'X)^{-1}K\theta = m$$

Hence, $\theta = [K'(X'X)^{-1}K]^{-1}(K\widehat{b_c} - m)$ (6)

Thus, the constrained least-square estimator

$$\widehat{b_c} = \hat{b} - (X'X)^{-1}K[K'(X'X)^{-1}K]^{-1}(K\widehat{b_c} - m) \qquad (7)$$

We have estimated b under the null hypothesis $H: K'b = m$. We now show that the corresponding residual sum of squares is SSE+Q where Q is the numerator sum of squares of the $F -$ Statistic used in equation (3), $F(H)$. We consider the residual:

$$\left(y - X\widehat{b_c}\right)'\left(y - X\widehat{b_c}\right) = \left[y - X\hat{b} + X\left(\hat{b} - \widehat{b_c}\right)\right]'\left[y - X\hat{b} + X\left(\hat{b} - \widehat{b_c}\right)\right]$$

$$= \left(y - X\hat{b}\right)'\left(y - X\hat{b}\right) + \left(\hat{b} - \widehat{b_c}\right)'X'\left(y - X\hat{b}\right) + \left(y - X\hat{b}\right)'X\left(\hat{b} - \widehat{b_c}\right)' + \left(\hat{b} - \widehat{b_c}\right)'X'X\left(\hat{b} - \widehat{b_c}\right)$$

$$= \left(y - X\hat{b}\right)'\left(y - X\hat{b}\right) + \left(\hat{b} - \widehat{b_c}\right)'X'X\left(\hat{b} - \widehat{b_c}\right) \qquad (8)$$

Since, $\{ x'\left(y - X\hat{b}\right) = X'y - X'X(X'X)^{-1}X'y = 0\}$

Substituting the constrained least-square estimator equation (7) into equation (8), we get:

$$\left(y - X\widehat{b_c}\right)'\left(y - X\widehat{b_c}\right) = SSE + \left(K\widehat{b_c} - m\right)'[K'(X'X)^{-1}K]^{-1}K'(X'X)^{-1}X'X(X'X)^{-1}K[K'(X'X)^{-1}K]^{-1}(K\widehat{b_c} - m)'$$

$$= SSE + (K\widehat{b_c} - m)'[K'(X'X)^{-1}K]^{-1}(K\widehat{b_c} - m) \qquad (9)$$

$$= SSE + Q$$

In deriving the constrained least-square estimator, we used an exact constraint $K'b = m$.

Four Common Hypothesis

i) First consider $H: b = 0$. The test of this hypothesis has already been considered in the analysis of variance tables. However, it illustrates the reduction of F(H) to the F-statistic of the analysis of variance tables. To apply $F(H)$ we need to specify K′ and m for the equation $K'b = m$. To apply $F(H)$we need to specify K′ and m. We have that $K' = I, s = k + 1 \; and \; m = 0$. Thus, $[K'(X'X)^{-1}K]^{-1}$ becomes $X'X$. Then, as before,

$$F(H) = \frac{\hat{b}X'X\hat{b}}{(k+1)\widehat{\sigma^2}} = \frac{SSR}{r} * \frac{N-r}{SSE}$$

Under the null hypothesis $F(R) \sim F_{(r,N-r)}$ where $r = k + 1$

The corresponding value of $\widehat{b_c} = \hat{b} - (X'X)^{-1}[(X'X)^{-1}]^{-1}\hat{b} = 0$

ii) We now consider $H: b = b_0$,that is $b_i = b_{(i0)}$ for all i. Rewriting $b = b_0$ as $K'b = m$ gives: $K' = I, s = k + 1, m = b_0 \; and \; [K'(X'X)^{-1}K]^{-1} = X'X$. Thus,

$$F(H) = \frac{(\hat{b}-b_0)\prime X'X\widehat{(b}-b_0)}{(k+1)\widehat{\sigma^2}} \qquad (10)$$

Under the null hypothesis is:

$$\widehat{b_c} = \hat{b} - [(X'X)^{-1}]^{-1}(\hat{b} - b_0) = b_0$$

iii) Now, consider $H: \lambda'b = m$. in this case, we have $K' = \lambda', s = 1 \; and \; m = m$. Since λ' is a vector,

$$F(H) = \frac{(\lambda'\hat{b}-m)' \, [\lambda'(X'X)^{-1}\lambda]^{-1}(\lambda'\hat{b}-m)}{\widehat{\sigma^2}} = \frac{(\lambda'\hat{b}-m)^2}{\lambda'(X'X)^{-1}\lambda\widehat{\sigma^2}}$$

Under the null hypothesis,$F(H)$ has the $F_{(1,N-r)}$-distribution.

Hence, $\sqrt{F(H)} = \frac{(\lambda'\hat{b}-m)}{\hat{\sigma}\sqrt{\lambda'(X'X)^{-1}\lambda}} \sim t_{(N-r)}$

This is as one would expect because $\lambda'\hat{b}$ is normally distributed with variance $\lambda'(X'X)^{-1}\lambda$

For this hypothesis, the value of $\widehat{b_c}$ is

$$\widehat{b_c} = \hat{b} - (X'X)^{-1}\lambda[\lambda'(X'X)^{-1}\lambda]^{-1}(\lambda'\hat{b} - m)$$

$$= \hat{b} - \frac{(\lambda'\hat{b}-m)}{\lambda'(X'X)^{-1}\lambda}(X'X)^{-1}\lambda$$

Observe that:

$$\lambda'\widehat{b_c} = \lambda'\hat{b} - \lambda'(X'X)^{-1}\lambda[\lambda'(X'X)^{-1}\lambda]^{-1}(\lambda'\hat{b} - m)$$

$$= \lambda'\hat{b} - (\lambda'\hat{b} - m) = m$$

Thus, $\widehat{b_c}$ satisfies the null hypothesis $H: \lambda'b = m$

Note: At this point, it is appropriate to comment on the lack of emphasis being given to the t-test in hypothesis testing. The equivalence of t-statistics with F-statistics with one degree of freedom in the numerator makes it unnecessary to consider t-tests. Whenever a t-test might be proposed, the hypothesis to be tested can be put in the form H: λ' b = m and the F-statistic F(H) derived as here. If the t-statistic is insisted upon, it is then obtained as √F(H). No further discussion of using the t-test is therefore necessary.

iv) We now consider the case where the null hypothesis is that the first q coordinate of b is zero, that is, $H: b_q = 0$ i.e., $b_i = 0$ for $i = 0, 1, 2, \ldots q - 1$, for $q < k$. In this case, we have $K' = [I_q \quad 0]$ and $m = 0$ so that $s = q$. We write

$\widehat{b_q'} = [b_0 b_1 \ldots \ldots \ldots . . b_{(q-1)}]$

and partition $b, \hat{b}$ *and* $(X'X)^{-1}$ accordingly. Thus,

$$b = \begin{pmatrix} b_q \\ b_p \end{pmatrix} \qquad \hat{b} = \begin{pmatrix} \hat{b}_q \\ \hat{b}_p \end{pmatrix} \qquad \text{And, } (X'X)^{-1} = \begin{pmatrix} Tqq & Tqp \\ Tpq & Tpp \end{pmatrix}$$

Where $p + q =$ the order of $b = k + 1$.then in $F(H)$ in general hypothesis:$K'\hat{b} = \widehat{b_q}$

And, $[K'(X'X)^{-1}K]^{-1} = {T_{(qq)}}^{-1}$

Giving $F(H) = \frac{\widehat{b_q^T}\, {T_{(qq)}}^{-1}\, \widehat{b_q}}{q\widehat{\sigma^2}}$ (11)

In the numerator, we recognize the result of "invert part of the inverse". That means, take the inverse of $X'X$ and invert that part of it that corresponds to b_q of the hypothesis $H: b_q = 0$. Although demonstrated here that for a b_q that consists of the first $qb's$ in b, it clearly applies to any subset of q b's. In particular, for just one b, it leads to the usual F-test on one degree of freedom, equivalent to t −test.

The estimator of b under this hypothesis is:

$$\hat{b}_c = \hat{b} - (X'X)^{-1}\begin{bmatrix} I_q \\ 0 \end{bmatrix} T_{qq}^{-1}(\hat{b}_q - 0)$$

$$= \hat{b} - \begin{bmatrix} T_{qq} \\ T_{pq} \end{bmatrix} T_{qq}^{-1}(\hat{b}_q) = \begin{bmatrix} \hat{b}_q \\ \hat{b}_p \end{bmatrix} - \begin{bmatrix} \hat{b}_q \\ T_{pq}T_{qq}^{-1}\hat{b}_q \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ \hat{b}_p - T_{pq}T_{qq}^{-1}\hat{b}_q \end{bmatrix}$$

Thus, the estimators of b's or not in the hypothesis are $\widehat{b_p} - T_{(pq)T_{(qq)}}{}^{-1}\widehat{b_q}$.

The expression obtained for $F(H)$ and $\widehat{b_c}$ for these four-hypothesis concerning b are in term of $\hat{b}$.

1.8 Self – Assessment Exercise

1. The deciles of a normal distribution are:

$d_1 = 17.5056$	$d_4 = 20.6764$	$d_7 = 23.992$
$d_2 = 18.7189$	$d_5 = 21.6681$	$d_8 = 25.5026$
$d_3 = 19.7684$	$d_6 = 22.7592$	$d_9 = 27.8952$

   Estimate by the method of least squares, the mean and standard deviation of the distribution.

2. For the model $E(y) = \eta = bI$, $V(y) = \sigma^2 I$,describe the estimation space and error space and find the least square estimate for b. Show that X and $y - \hat{\eta}$ are orthogonal. Also find $E(\hat{b})$.

3. For the model $= Xb + e, e \sim N(0, \sigma^2 I)$, $g(y)$ is some function of y, such that its expected value is identically equal to zero. Show that the covariance between g(y) and the element to $X'y$ is null.

   Let $L(y)$ be any function of y, such that its expected value is $\lambda' b$. Let $\lambda'\hat{b}$ is the BLUE of $\lambda' b$. Defining $g(y) = L(y) - \lambda' b$, show that $V(L(y)) > V(\lambda'\hat{b})$.

4. Suppose $E(y_1) = E(y_2) = \theta$,but $V(y_1) = 5\sigma^2, cov(y_1, y_2) = \sigma^2, V(y_2) = 2\sigma^2$.Show that the BLUE of θ is $\widehat{\theta'} = (y_1 + 4y_2)/5$

5. When y has the variance covariance matrix V ,prove that the covariance of the BLUE's of $p'b$ and $q'b$ is $p'(X'V^{-1}X)^{-1}q$.

1.9 Summary

The unit covers the basic concepts of linear estimation technique of model building, estimable functions etc. In this unit, the procedure of obtaining the Best Linear Unbiased Estimator (BLUE) is discussed in detail. Also, the Markov theorem distribution of quadratic form is explained. The generalized F and t tests are also covered, which are used to test the general hypotheses of linear estimation.

1.10 References

- Gujarathi, D. (1979). *Basic Econometrics*, McGraw hill.
- Rao, C.R. (1973). *Linear statistical inference and its applications*. Wiley Eastern.
- Searle, S.R. and Gruber M.H.J. (2017). *Linear Models*, Second Edition. Wiley.

1.11 Further Reading

- Bapat, R. (2012). *Linear Algebra and Linear Models*. Springer, Germany.
- Kshirsagar, A.M. (1983). *A Course in Linear Models*. Taylor & Francis.
- Montgomery, D.C., Peck, E.A., Vining, G.G. (2003). *Introduction to Linear Regression Analysis*, John Wiley.
- Rencher, A.C. and Schaalje, G.B. (2008). *Linear Models in Statistics*. Wiley.

UNIT:2 ANALYSIS OF VARIANCE- I

Structure

2.1 Introduction

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. For example, when 5 different fertilizers are applied to four plots each, then we may be interested in finding whether the fertilizers have any significant effect on the yield. In other words, we want to see whether the samples are coming from the same normal population.

In any set of observations, the variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes, but mainly classified as:

(i) ***Assignable cause of variation:*** The assignable cause of variation can be identified, measured and controlled.

(ii) ***Chance cause of variation:*** The chance cause of variation is beyond the control of human hand and cannot be traced separately.

Analysis of Variance (ANOVA) consists of estimation of the amount of effects due to each of independent factors (causes) separately and compare the estimates of effects due to assignable factors (causes) with estimates of the effects due to chance factor (cause) or experimental error or simple error.

The following assumptions are made in any analysis of variance procedure:

(1) The observations are independent.

(2) Parent population from which observations are taken is normal; and

(3) Various treatment and environmental effects are additive in nature.

2.2 Objectives

After going through this unit, you should be able to:

- Acquire the knowledge analysis of variance (ANOVA) concept,
- Perform the analysis of variance in one-way classified data with equal (one) observation per cell,
- Able to analyze the two-way classified data with equal (one) observation per cell using ANOVA method.

2.3 Analysis of Variance: One-Way Classification with One Observation Per Cell

Suppose there are n observations y_{ij},$(i = 1,2,\cdots,k; j = 1,2,\cdots,n_i)$ of a random variable *Y* are grouped into *k* groups of size $n_1, n_2, \cdots, n_k$ respectively. Then *n* = $\sum_{i=1}^{k} n_i$ and the observation table is as follows:

Groups	***Observations***	***Total***	***Mean***
1	y_{11} y_{12} ……………… y_{1n_1}	$T_{1.} = \sum_{j=1}^{n_1} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{n_1}$
2	y_{21} y_{22} ……………… y_{2n_2}	$T_{2.} = \sum_{j=1}^{n_2} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{n_2}$
⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
i	y_{i1} y_{i2} ……………… y_{in_i}	$T_{i.} = \sum_{j=1}^{n_i} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{n_i}$

⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
k	y_{k1} y_{k2} ……………… y_{kn_k}	$T_{k.} = \sum_{j=1}^{n_k} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{n_k}$
Total		$T_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$	$\bar{y}_{..} = \frac{T_{..}}{n}$

2.3.1 Statistical Analysis

The total variation in the observation can be split into the following two components.

(i) The variation between the classes or *assignable cause* of variation and

(ii) The variation within the classes or *chance cause* of variation.

Hence the mathematical model is given by:

$y_{ij} = \mu_i + e_{ij}; j = 1,2,\cdots,n_i, i = 1,2,\cdots,k,$

where μ_i is the average effect of the i^{th}group, which can be split as:

$\mu_i = \mu + \mu_i - \mu = \mu + \alpha_i$ with $\alpha_i = \mu_i - \mu, i = 1,2,\cdots,k$ and $\mu = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i$.

Hence,

$y_{ij} = \mu + \alpha_i + e_{ij}; j = 1,2,\cdots,n_i, i = 1,2,\cdots,k;$ (1)

Where:

y_{ij} is the j^{th} observation of i^{th} class; $j = 1,2,\cdots,n_i, i = 1,2,\cdots,k,$

μ is the general mean effect,

α_i is the additive effect due to i^{th}group and

e_{ij} is the error effect due to chance and these are assumed to be iid random variables each following $N(0,\sigma_e^2); j = 1,2,\cdots,n_i, i = 1,2,\cdots,k.$

The side condition is $\sum_{i=1}^{k} n_i\alpha_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = n\mu - n\mu = 0.$

Assumptions

The statistical analysis of this layout is based on the following assumptions.

(i) All the observations are mutually independent.

(ii) Different effects are additive in nature.

(iii) e_{ij}'s are *iid* random variables each following $N(0,\sigma_e^2); j = 1,2,\cdots,n_i, i = 1,2,\cdots,k.$

The null hypothesis to be tested is:

H_0: The groups do not differ significantly or there is no additive effect due to different groups. In other words, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

Summing (1) over j and dividing by n_i, we get

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k, \qquad (2)$$

where $\bar{e}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/n_i)$.

Summing (1) over i and j and dividing by n, we get:

$$\bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \mu + \bar{e}_{..} = \mu + \bar{e}_{..}, \qquad (3)$$

where $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are *iid* random variables each distributed as $N(0, \sigma_e^2/n)$.

Now the total variation in each observation is given by the total sum of squares as:

$$\text{T.S.S.} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_{i.} - \bar{y}_{..} + y_{ij} - \bar{y}_{i.}\right)^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

$$= \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2.$$

Or T.S.S. = S.S.G. + S.S.E

Where,

T.S.S = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$;

S.S.G = Sum of squares due to groups = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$; and

S.S.E = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$.

Degrees of Freedom for various Sums of Squares

T.S.S = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$ is computed from n quantities of the form $\left(y_{ij} - \bar{y}_{..}\right)$ with one constraint $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right) = 0$. Hence, T.S.S will have n – 1 degrees of freedom.

S.S.G = Sum of squares due to groups = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ is computed from k quantities of the form $(\bar{y}_{i.} - \bar{y}_{..})$ with one constraint $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Hence, S.S.G will have k – 1 degrees of freedom.

S.S.E = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$ is computed from n quantities of the form $\left(y_{ij} - \bar{y}_{i.}\right)$ with k constraints $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right) = 0$. Hence, S.S.E will have n – k degrees of freedom.

Mean Sum of Squares

The sum of squares divided by its degrees of freedom gives the corresponding mean sum of squares. Thus:

Mean sum of squares due to groups = M.S.G.= $\frac{S.S.G.}{k-1}$.

Mean sum of squares due to error = M.S.E.= $\frac{S.S.E.}{n-k}$

2.3.2 Least Square Estimates

In the mathematical model (1), μ and $\alpha_i, i = 1,2,\cdots,k$ are the unknown parameters which have to be estimated by the principle of least squares. Hence, we consider the sum of squares due to errors, which is given by:

$$S.S.E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right)^2 \qquad (4)$$

Differentiating (4) with respect to μ and α_i and equating to zero individually, we get:

$$\frac{dS.S.E}{d\mu} = 0 \Rightarrow -2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = n\mu + \sum_{i=1}^{k} n_i\alpha_i = n\mu \ \ [\because \sum_{i=1}^{k} n_i\alpha_i = 0 \text{ by side condition.}]$$

Hence, the estimate of μ is given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}.$$

$$\frac{dS.S.E}{d\alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0,\ i = 1,2,\cdots,k.$$

$$\Rightarrow \sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$\Rightarrow \sum_{j=1}^{n_i} y_{ij} = n_i\mu + n_i\alpha_i$

$\Rightarrow \hat{\alpha}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$.

Variance of the Estimates

We have $\hat{\mu} = \bar{y}_{..}$ and $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$.

$V(\hat{\mu}) = E[\bar{y}_{..} - E(\bar{y}_{..} - \alpha_0)]^2 = E[\mu + \bar{e}_{..} - \mu]^2 = E[\bar{e}_{..}]^2 = E(\bar{e}_{..}^2)$

$= V(\bar{e}_{..}) = \frac{\sigma_e^2}{n}$.

Also, we have $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{y}_{i.} - \bar{y}_{..} - E(\bar{y}_{i.} - \bar{y}_{..})$

$= \mu + \alpha_i + \bar{e}_{i.} - \mu - \alpha_0 - \bar{e}_{..} - E(\mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..})$

$= \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} - \mu + \mu$

$= \alpha_i + \bar{e}_{i.} - \bar{e}_{..}$.

Hence, $V(\hat{\alpha}_i) = E[\alpha_i + \bar{e}_{i.} - \bar{e}_{..}]^2 = E(\alpha_i^2) + E[\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}]$

$= \alpha_i^2 + E(\bar{e}_{i.}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{i.}\bar{e}_{..})$.

Now, $E(\bar{e}_{i.}\bar{e}_{..}) = E(\frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij} \frac{1}{kn_i}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij})$

$= \frac{1}{kn_i^2} E\left[e_{i1}^2 + e_{i2}^2 + \cdots + e_{in_i}^2\right] + \frac{1}{kn_i^2} E\left[\sum_{j=1}^{n_i} e_{ij} \sum_{h\neq i=1}^{k}(e_{h1} + \cdots + e_{hn_i})\right]$

$= \frac{1}{kn_i^2} E\left[e_{i1}^2 + e_{i2}^2 + \cdots + e_{in_i}^2\right]$ since $E(e_{ij}e_{hj}) = 0$ for $h \neq i$;

$= \frac{1}{kn_i^2}\sum_{j=1}^{n_i} E\left(e_{ij}^2\right) = \frac{1}{kn_i^2}\sum_{j=1}^{n_i} V\left(e_{ij}\right)$

$= \frac{1}{kn_i^2} n_i\sigma_e^2 = \frac{\sigma_e^2}{kn_i}$.

Hence, $V(\hat{\alpha}_i) = \alpha_i^2 + \frac{\sigma_e^2}{n_i} + \frac{\sigma_e^2}{n} - 2\frac{\sigma_e^2}{kn_i} = \alpha_i^2 + \frac{\sigma_e^2}{n_i}\left(1 - \frac{2}{k}\right) + \frac{\sigma_e^2}{n}$.

In particular if all group sizes are equal, say to r, i.e., if $n_i = r, \forall\, i = 1,2,\cdots,k$, then n = rk and:

$V(\hat{\alpha}_i) = \alpha_i^2 + \frac{\sigma_e^2}{r}\left(1 - \frac{2}{k}\right) + \frac{\sigma_e^2}{rk} = \alpha_i^2 + \frac{\sigma_e^2}{r}\left(1 - \frac{2}{k} + \frac{1}{k}\right) = \alpha_i^2 + \frac{(k-1)\sigma_e^2}{rk}$.

Expectation of Sum of Squares

We have $y_{ij} = \mu + \alpha_i + e_{ij}; j = 1,2,\cdots,n_i, i = 1,2,\cdots,k$;

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mu + \alpha_i + \bar{e}_{i.},\ \forall\ i = 1,2,\cdots,k, \text{ and}$$

$$\bar{y}_{..} = \mu + \bar{e}_{..},$$

Then:

$$E(y_{ij}^2) = E(\mu^2 + \alpha_i^2 + e_{ij}^2 + 2\mu\alpha_i + 2\mu e_{ij} + 2\alpha_i e_{ij})$$

$$= E(\mu^2) + E(\alpha_i^2) + E(e_{ij}^2) + 2\mu E(\alpha_i) + 2\mu E(e_{ij}) + 2E(\alpha_i)E(e_{ij})$$

$$= \mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i .$$

$$E(\bar{y}_{i.}^2) = E(\mu^2 + \alpha_i^2 + \bar{e}_{i.}^2 + 2\mu\alpha_i + 2\mu\bar{e}_{i.} + 2\alpha_i\bar{e}_{i.})$$

$$= E(\mu^2) + E(\alpha_i^2) + E(\bar{e}_{i.}^2) + 2\mu E(\alpha_i) + 2\mu E(\bar{e}_{i.}) + 2E(\alpha_i)E(\bar{e}_{i.})$$

$$= \mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i.$$

$$E(\bar{y}_{..}^2) = E(\mu^2 + \bar{e}_{..}^2 + 2\mu\bar{e}_{..})$$

$$= E(\mu^2) + E(\bar{e}_{..}^2) + 2\mu E(\bar{e}_{..}) = \mu^2 + \frac{\sigma_e^2}{n}.$$

$$E(S.S.G.) = E\{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2\}$$

$$= E\{\sum_{i=1}^{k} n_i\bar{y}_{i.}^2 - n\bar{y}_{..}^2\}$$

$$= \sum_{i=1}^{k} n_i E(\bar{y}_{i.}^2) - nE(\bar{y}_{..}^2)$$

$$= \sum_{i=1}^{k} n_i\left(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i\right) - n(\mu^2 + \frac{\sigma_e^2}{n})$$

$$= n\mu^2 + \sum_{i=1}^{k} n_i\,\alpha_i^2 + k\,\sigma_e^2 + 2\mu\sum_{i=1}^{k} n_i\alpha_i - n\mu^2 - \sigma_e^2$$

$$= \sum_{i=1}^{k} n_i\,\alpha_i^2 + (k-1)\sigma_e^2.$$

Or $E(M.S.G.) = E\left(\frac{S.S.G}{k-1}\right) = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i\,\alpha_i^2 + \sigma_e^2.$

Now $E(S.S.E.) = E\{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\}$

$$= E\{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} n_i\bar{y}_{i.}^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} E(y_{ij}^2) - \sum_{i=1}^{k} n_i E(\bar{y}_{i.}^2)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i) - \sum_{i=1}^{k} n_i(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i)$$

$= n\mu^2 + \sum_{i=1}^{k} n_i \alpha_i^2 + n\sigma_e^2 + 2\mu \sum_{i=1}^{k} n_i\alpha_i - n\mu^2 - \sum_{i=1}^{k} n_i \alpha_i^2 - k\sigma_e^2 -$

$2\mu \sum_{i=1}^{k} n_i\alpha_i$

$= (n - k)\sigma_e^2$

Or $E(M.S.E.) = E\left(\frac{S.S.E}{n-k}\right) = \sigma_e^2$

Thus, under H_0, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. Hence,

$E(M.S.G.) = \sigma_e^2 = E(M.S.E.)$.

Also, under H_0, S.S.G. follows a χ^2 distribution with k – 1 degrees of freedom and S.S.E. follows a χ^2 distribution with n – k degrees of freedom.

Hence, for testing H_0, the test statistic is given by $F = \frac{S.S.G/(k-1)}{S.S.E./(n-k)} = \frac{M.S.G}{M.S.E}$ which will follow a central F distribution with k – 1 and n – k degrees of freedom.

2.2.3 ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Groups	k – 1	$S.S.G. = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$M.S.G = \frac{S.S.G.}{k-1}$	$F = \frac{M.S.G.}{M.S.E}$
Error	n – k	$S.S.E. = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$	$M.S.E. = \frac{S.S.E}{n-k}$	
Total	n – 1	$T.S.S. = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2$		

If the calculated value of *F* is greater than the tabulated value of *F* at *k* – 1 and *n* – *k* degrees of freedom, then reject the null hypothesis H_0 otherwise it may be accepted.

Critical Difference

If the null hypothesis is rejected, then we may test for the equality of two classes means i.e.

H_0, $\mu_i = \mu_{i'}$; i≠i'=1,2…p

Here we apply t-test satisfying the test statistic t:

$$|t| = \frac{\overline{y_{i.}} - \overline{y_{i.'}}}{\sqrt{MSE\left(\frac{1}{ni} + \frac{1}{ni'}\right)}} \sim t_{\frac{\alpha}{2}}(n - k)$$

if $n_1 = n_2 = \cdots \dots . n_p = n_0$, then

$|t| = \frac{\overline{y_{i.}} - \overline{y_{i.'}}}{\sqrt{\frac{2MSE}{n_0}}} \sim \quad t_{\frac{\alpha}{2}}(n-k)$

and if $|t| \leq t_{\frac{\alpha}{2}}(n-k)$, then we accept our null hypothesis H_0 at α x 100% level of significance, otherwise reject.

The quantity $\sqrt{\frac{2MSE}{n_0}}$, $t_{\frac{\alpha}{2}}(n-k)$ is known as the "critical difference" or "least significant difference".

For Practical Calculations

We have T.S.S = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}^2 + \bar{y}_{..}^{\,2} - 2y_{ij}\bar{y}_{..})$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 + n\bar{y}_{..}^{\,2} - 2\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 + n\bar{y}_{..}^{\,2} - 2n\bar{y}_{..}^{\,2}$

$[\because \bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \Rightarrow n\bar{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}]$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}_{..}^{\,2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{n}$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T_{..}^{\,2}}{n}$

T.S.S = Raw Sum of Squares (R.S.S.) – Correction factor (C.F.).

S.S.G $= \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i.}^{\,2} + \bar{y}_{..}^{\,2} - 2\bar{y}_{i.}\,\bar{y}_{..})$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} + n\,\bar{y}_{..}^{\,2} - 2\,\bar{y}_{..}\sum_{i=1}^{k} n_i\,\bar{y}_{i.}$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} + n\,\bar{y}_{..}^{\,2} - 2n\bar{y}_{..}^{\,2}$

$= \sum_{i=1}^{k} n_i\bar{y}_{i.}^{\,2} - n\,\bar{y}_{..}^{\,2} = \sum_{i=1}^{k} n_i\left(\frac{T_{i.}}{n_i}\right)^2 - n\left(\frac{T_{..}}{n}\right)^2$

$= \sum_{i=1}^{k}\frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{n} = \sum_{i=1}^{k}\frac{T_{i.}^2}{n_i} - \text{C.F.}$

S.S.E. = T.S.S. – S.S.G.

Example: To assess the significance of possible variation in performance in a certain test between the convent schools of a city, a common test was given to a number of students taken at random from the senior fifth class of each of the four schools concerned. The results are given below. Make an analysis of variance of data.

Schools

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution:

Sample-I		*Sample-II*		*Sample-III*		*Sample-IV*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\sum X_1 = 45$	$\sum X_1^2 = 421$	$\sum X_2 = 50$	$\sum X_2^2 = 558$	$\sum X_3 = 60$	$\sum X_3^2 = 824$	$\sum X_4 = 65$	$\sum X_4^2 = 875$

$G = \sum X_1 + \sum X_2 + \sum X_3 + \sum X_4 = 220$

Correction Factor $= \frac{T^2}{N} = \frac{(220)^2}{20} = 2420$

Total Sum of Squares (TSS) $= (\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2) - \frac{T^2}{N} = 2678 - 2420 = 258$

Sum of Squares between Groups (SSG) $= \frac{(\sum X_1)^2}{N} + \frac{(\sum X_2)^2}{N} + \frac{(\sum X_3)^2}{N} + \frac{(\sum X_4)^2}{N} - \frac{T^2}{N}$

$= 2470 - 2420 = 50$

Sum of Squares due to Error (SSE) = Total Sum of Squares – Sum of Squares between samples

= 258 – 50 = 208

ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$F_{Cal.}$	***$F_{Tab.}$***
Groups	3	50	16.7	1.285	$F_{(3,5)} = 3.24$

Error	16	208	13.0		
Total	19	258			

The calculated value of F is less than the tabulated value and hence, the difference in the mean value of the samples is not significant i.e., the samples could have come from the same universe.

2.4 Analysis of Variance: Two-Way Classification with One Observation Per Cell

Suppose there are *n* observations y_{ij},$(i = 1,2,\cdots,k; j = 1,2,\cdots,h)$ of a random variable *Y* are grouped into *k* rows and *h* columns respectively. Then *n* = hk and the observation table is as follows:

Rows	***Columns***						***Row Totals***	***Row Means***
	1	***2***	$\cdots$	***j***	$\cdots$	***h***		
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1h}	$T_{1.} = \sum_{j=1}^{h} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{h}$
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2h}	$T_{2.} = \sum_{j=1}^{h} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{h}$
$\vdots$ $\vdots$				$\vdots$ $\vdots$			$\vdots$ $\vdots$	$\vdots$ $\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ih}	$T_{i.} = \sum_{j=1}^{h} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{h}$
$\vdots$ $\vdots$				$\vdots$ $\vdots$			$\vdots$ $\vdots$	$\vdots$ $\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kh}	$T_{k.} = \sum_{j=1}^{h} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{h}$
Column Totals	$T_{.1} = \sum_{i=1}^{k} y_{i1}$ $T_{.2} = \sum_{i=1}^{k} y_{i2} \cdots T_{.j} = \sum_{i=1}^{k} y_{ij} \cdots$ $T_{.h} = \sum_{i=1}^{k} y_{ih}$						$T_{..} = \sum_{i=1}^{k} \sum_{j=1}^{h} y_{ij}$	

Column Means	$\bar{y}_{.1} = \frac{T_{.1}}{k}$ $\bar{y}_{.2} = \frac{T_{.2}}{k}$ $\cdots$ $\bar{y}_{.j} = \frac{T_{.j}}{k}$ $\cdots$ $\bar{y}_{.h} = \frac{T_{.h}}{k}$		$\bar{y}_{..} = \frac{T_{..}}{hk}$

2.4.1 Statistical Analysis

The total variation in the observation can be split into the following two components:

(i) The variation between the classes or *assignable cause* of variation which are due to classification into different rows and column, and

(ii) the variation within the rows or columns or *chance cause* of variation.

Let y_{ij} denote the value of the observation in the (i, j)th cell and suppose that y_{ij}'s are iid random variables, distributed according to $N(\mu_{ij}, \sigma_e^2)$. Then the mathematical model is:

$y_{ij} = \mu_{ij} + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h$

where $e_{ij}'s$ are the error effect due to chance and these are assumed to be iid random variables each following $N(0, \sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h.$

μ_{ij} is further split into:

(1) $\mu = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij}$, the over all mean;

(2) the row effect $\alpha_i = \mu_{i.} - \mu$, where $\mu_{i.} = \frac{1}{h}\sum_{j=1}^{h}\mu_{ij}$; and

(3) the column effect $\beta_j = \mu_{.j} - \mu$, where $\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}$.

Obviously,

$\sum_{i=1}^{k}\alpha_i = \sum_{i=1}^{k}(\mu_{i.} - \mu) = \sum_{i=1}^{k}\mu_{i.} - k\,\mu = k\,\mu - k\,\mu = 0.$

Similarly,

$\sum_{j=1}^{h}\beta_j = \sum_{j=1}^{h}\left(\mu_{.j} - \mu\right) = \sum_{j=1}^{h}\mu_{.j} - h\,\mu = h\,\mu - h\,\mu = 0.$

Thus,

$\mu_{ij} = \mu + \mu_{i.} - \mu + \mu_{.j} - \mu = \mu + \alpha_i + \beta_j$

Hence the mathematical model is given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h \qquad (1)$$

Where,

y_{ij} is the observation of i^{th} row and j^{th} column; $i = 1,2,\cdots,k, j = 1,2,\cdots,h,$

μ is the general mean effect

α_i is the additive effect due to i^{th}row; $i = 1,2,\cdots,k;$

β_j is the additive effect due to j^{th}column; $i = 1,2,\cdots,k;$ and

$e_{ij}'s$ are the error effect due to chance and these are assumed to be iid random variables each following $N(0,\sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h.$

The side conditions are $\sum_{i=1}^{k}\alpha_i = \sum_{j=1}^{h}\beta_j = 0.$

Summing (1) over j and dividing by h, we get

$$\bar{y}_{i.} = \frac{1}{h}\sum_{j=1}^{h} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k, \tag{2}$$

and

$\bar{e}_{i.} = \frac{1}{h}\sum_{j=1}^{h} e_{ij}$ are iid random variables each distributed as $N(0,\sigma_e^2/h)$.

Summing (1) over i and dividing by k, we get

$$\bar{y}_{.j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij} = \mu + \beta_j + \bar{e}_{.j}, \forall\, j = 1,2,\cdots,h, \tag{3}$$

and

$\bar{e}_{.j} = \frac{1}{k}\sum_{i=1}^{k} e_{ij}$ are iid random variables each distributed as $N(0,\sigma_e^2/k)$.

Summing (1) over i and j and dividing by $n = hk$, we get

$$\bar{y}_{..} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij} = \mu + \bar{e}_{..}, \tag{4}$$

where $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are iid random variables each distributed as $N(0,\sigma_e^2/hk)$.

The null hypothesis to be tested is:

H_{01}: The rows do not differ significantly or there is no additive effect due to different rows. In other words, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ and

H_{02}: The columns do not differ significantly or there is no additive effect due to different columns. In other words, $\beta_1 = \beta_2 = \cdots = \beta_h = 0$.

Now the total variation in each observation is given by the total sum of squares as

T.S.S. = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(\bar{y}_{i.} - \bar{y}_{..} + \bar{y}_{.j} - \bar{y}_{..} + y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$

$= h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2.$

Or T.S.S. = S.S.R. + S.S.C. + S.S.E,

Where,

T.S.S = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$.

S.S.R = Sum of squares due to rows = $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$.

S.S.C = Sum of squares due to columns = $k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$ and

S.S.E = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$.

Degrees of freedom

T.S.S = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$ is computed from hk quantities of the type $\left(y_{ij} - \bar{y}_{..}\right)$ with one restriction that $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right) = 0$. Hence, it has hk – 1 degrees of freedom.

S.S.R = $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$ is computed from k quantities of the type $(\bar{y}_{i.} - \bar{y}_{..})$ with one restriction of the type $\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Therefore S.S.R has k – 1 degrees of freedom.

S.S.C = $k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$ is computed from h quantities of the type $\left(\bar{y}_{.j} - \bar{y}_{..}\right)$ with one restriction of the type $\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right) = 0$. Therefore S.S.R has h – 1 degrees of freedom.

Finally, S.S.E = $\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$= T.S.S – S.S.R. – S.S.C. Hence its degree of freedom is given by hk – 1 – (k – 1) – (h – 1) = hk –k –h + 1 = (h – 1). (k – 1).

2.4.2 Least Square Estimates

In the mathematical model (1), μ , α_i and β_j , $i = 1,2,\cdots,k, j = 1,2,\cdots,h$ are the unknown parameters which have to be estimated by the principle of least squares. Hence, we consider the sum of squares due to errors, which is given by:

S.S.E = $\sum_{i=1}^{k}\sum_{j=1}^{h}e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)^2$. (5)

Differentiating (5) with respect to μ, α_i and β_j and equating to zero individually, we get:

$$\frac{dS.S.E}{d\mu} = 0 \Rightarrow -2\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = hk\mu + h\sum_{i=1}^{k}\alpha_i + k\sum_{j=1}^{h}\beta_j = hk\mu \,.$$

Hence, the estimate of μ is given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}.$$

$$\frac{dS.S.E}{d\alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0,\ i = 1,2,\cdots,k.$$

$$\Rightarrow \sum_{j=1}^{h}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{j=1}^{h} y_{ij} = h\mu + h\alpha_i + \sum_{j=1}^{h}\beta_j$$

$$\Rightarrow \hat{\alpha}_i = \frac{1}{h}\sum_{j=1}^{h} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..} = \bar{y}_{i.} - \bar{y}_{..}$$

$$\frac{dS.S.E}{d\beta_j} = 0 \Rightarrow -2\sum_{i=1}^{k}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0,\ j = 1,2,\cdots,h.$$

$$\Rightarrow \sum_{i=1}^{k}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k} y_{ij} = k\mu + \sum_{i=1}^{k}\alpha_i + k\beta_j = k\mu + k\beta_j$$

$$\Rightarrow \hat{\beta}_j = \frac{1}{k}\sum_{i=1}^{k} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..} = \bar{y}_{.j} - \bar{y}_{..}$$

Variance of the estimates

We have $\hat{\mu} = \bar{y}_{..}$, $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ and $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$

$$V(\hat{\mu}) = E[\bar{y}_{..} - E(\bar{y}_{..})]^2 = E[\mu + \bar{e}_{..} - \mu]^2$$

$$= E\left[\bar{e}_{...}\right]^2 = V(\bar{e}_{..}) = \frac{\sigma_e^2}{hk}.$$

Also, we have $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{y}_{i.} - \bar{y}_{..} - E(\bar{y}_{i.} - \bar{y}_{..})$.

Now $\bar{y}_{i.} - \bar{y}_{..} = \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} = \bar{e}_{i.} - \bar{e}_{..} + \alpha_i$

$E(\bar{y}_{i.} - \bar{y}_{..}) = \alpha_i$.

Hence, $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{e}_{i.} - \bar{e}_{..} + \alpha_i - \alpha_i = \bar{e}_{i.} - \bar{e}_{..}$

Hence, $V(\hat{\alpha}_i) = E[\bar{e}_{i.} - \bar{e}_{..}]^2 = E[\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}]$

$$= E(\bar{e}_{i.}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{i.}\bar{e}_{..}).$$

Now, $E(\bar{e}_{i.}\bar{e}_{..}) = E(\frac{1}{h}\sum_{j=1}^{h} e_{ij} \frac{1}{kh}\sum_{i=1}^{k}\sum_{j=1}^{h} e_{ij})$

$$= \frac{1}{kh^2} E\left[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ih}^2\right] + \frac{1}{kh^2} E\left[\sum_{j=1}^{h} e_{ij} \sum_{g\neq i=1}^{k}\left(e_{g1} + \cdots + e_{gh}\right)\right]$$

$$= \frac{1}{kh^2} E\left[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ih}^2\right] \text{ since } E(e_{ij}e_{gj}) = 0 \text{ for } g \neq i;$$

$$= \frac{1}{kh^2}\sum_{j=1}^{n_i} E\left(e_{ij}^2\right) = \frac{1}{kh^2}\sum_{j=1}^{h} V\left(e_{ij}\right) = \frac{1}{kh^2} h\sigma_e^2 = \frac{\sigma_e^2}{kh}.$$

Hence, $V(\hat{\alpha}_i) = \frac{\sigma_e^2}{h} + \frac{\sigma_e^2}{kh} - 2\frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{h} - \frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{h}\frac{(k-1)}{k}$.

Similarly, for $V(\hat{\beta}_j)$, we have:

$\hat{\beta}_j - E(\hat{\beta}_j) = \bar{y}_{.j} - \bar{y}_{..} - E(\bar{y}_{.j} - \bar{y}_{..})$

$\bar{y}_{.j} - \bar{y}_{..} = \mu + \beta_j + \bar{e}_{.j} - \mu - \bar{e}_{..} = \bar{e}_{.j} - \bar{e}_{..} + \beta_j$

$E(\bar{y}_{.j} - \bar{y}_{..}) = \beta_j$.

Hence,

$V(\hat{\beta}_j) = E(\bar{e}_{.j} - \bar{e}_{..})^2 = E[\bar{e}_{.j}^2 + \bar{e}_{..}^2 - 2\bar{e}_{.j}\bar{e}_{..}]$

$$= E(\bar{e}_{.j}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{.j}\bar{e}_{..})$$

Now, $E(\bar{e}_{.j}\bar{e}_{..}) = E(\frac{1}{k}\sum_{i=1}^{k} e_{ij} \frac{1}{kh}\sum_{i=1}^{k}\sum_{j=1}^{h} e_{ij})$

$$= \frac{1}{hk^2} E\left[e_{1j}^2 + e_{2j}^2 + \cdots + e_{kj}^2\right] + \frac{1}{kh^2} E\left[\sum_{i=1}^{k} e_{ij} \sum_{l\neq j=1}^{h}(e_{1l} + \cdots + e_{kl})\right]$$

$$= \frac{1}{hk^2} E\left[e_{1j}^2 + e_{2j}^2 + \cdots + e_{kj}^2\right] \text{ since } E(e_{ij}e_{il}) = 0 \text{ for } l \neq j;$$

$$= \frac{1}{hk^2}\sum_{i=1}^{k} E\left(e_{ij}^2\right) = \frac{1}{hk^2}\sum_{i=1}^{k} V\left(e_{ij}\right) = \frac{1}{hk^2} k\sigma_e^2 = \frac{\sigma_e^2}{kh}.$$

Hence, $V(\hat{\beta}_j) = \frac{\sigma_e^2}{k} + \frac{\sigma_e^2}{kh} - 2\frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{k} - \frac{\sigma_e^2}{kh} = \frac{\sigma_e^2}{k}\frac{(h-1)}{h}$

Expectation of Sum of Squares

We have $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h$

$\bar{y}_{i.} = \frac{1}{h}\sum_{j=1}^{h} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k,$

$\bar{y}_{.j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij} = \mu + \beta_j + \bar{e}_{.j}, \forall\, j = 1,2,\cdots,h,$ and

$\bar{y}_{..} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij} = \mu + \bar{e}_{..},$

where $\bar{e}_{i.} = \frac{1}{h}\sum_{j=1}^{h} e_{ij}$ are iid random variables each distributed as $N(0, \sigma_e^2/h)$, $\bar{e}_{.j} = \frac{1}{k}\sum_{i=1}^{k} e_{ij}$ are iid random variables each distributed as $N(0, \sigma_e^2/k)$ and $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are iid random variables each distributed as $N(0, \sigma_e^2/hk)$.

Then:

$E(y_{ij}^2) = E(\mu^2 + \alpha_i^2 + \beta_j^2 + e_{ij}^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\mu e_{ij} + 2\alpha_i\beta_j + 2\alpha_i e_{ij} + 2\beta_j e_{ij})$

$= E(\mu^2) + E(\alpha_i^2) + E\left(\beta_j^2\right) + E\left(e_{ij}^2\right) + 2\mu E(\alpha_i) + 2\mu E\left(\beta_j\right) + 2\mu E(e_{ij}) +$

$2E(\alpha_i)E(\beta_j) + 2E\left(\alpha_i\right)E(e_{ij}) + 2E(\beta_j)E(e_{ij})$

$= \mu^2 + \alpha_i^2 + \beta_j^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\alpha_i\beta_j + \sigma_e^2$.

$E(\bar{y}_{i.}^2) = E(\mu^2 + \alpha_i^2 + \bar{e}_{i.}^2 + 2\mu\alpha_i + 2\mu\bar{e}_{i.} + 2\alpha_i\bar{e}_{i.})$

$= E(\mu^2) + E(\alpha_i^2) + E(\bar{e}_{i.}^2) + 2\mu E(\alpha_i) + 2\mu E(\bar{e}_{i.}) + 2E(\alpha_i)E(\bar{e}_{i.})$

$= \mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{h} + 2\mu\alpha_i.$

$E(\bar{y}_{.j}^2) = E(\mu^2 + \beta_j^2 + \bar{e}_{.j}^2 + 2\mu\beta_j + 2\mu\bar{e}_{.j} + 2\beta_j\bar{e}_{.j})$

$= E(\mu^2) + E(\beta_j^2) + E\left(\bar{e}_{.j}^2\right) + 2\mu E(\beta_j) + 2\mu E(\bar{e}_{.j}) + 2E(\beta_j)E(\bar{e}_{.j})$

$= \mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j.$

$E(\bar{y}_{..}^2) = E(\mu^2 + \bar{e}_{..}^2 + 2\mu\bar{e}_{..})$

$= E(\mu^2) + E(\bar{e}_{..}^2) + 2\mu E(\bar{e}_{..}) = \mu^2 + \frac{\sigma_e^2}{hk}.$

$E(S.S.R.) = E\{h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2\}$

$= E\{h\sum_{i=1}^{k}\bar{y}_{i.}^2 - hk\bar{y}_{..}^2\}$

$= h\sum_{i=1}^{k} E(\bar{y}_{i.}^2) - hkE(\bar{y}_{..}^2)$

$= h\sum_{i=1}^{k}\left(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{h} + 2\mu\alpha_i\right) - hk(\mu^2 + \frac{\sigma_e^2}{hk})$

$= hk\mu^2 + h\sum_{i=1}^{k}\alpha_i^2 + k\,\sigma_e^2 + 2\mu\sum_{i=1}^{k}\alpha_i - hk\mu^2 - \sigma_e^2$

$= h\sum_{i=1}^{k}\alpha_i^2 + (k-1)\sigma_e^2.$ [since $\sum_{i=1}^{k}\alpha_i = 0$].

Or $E(M.S.R.) = E\left(\frac{S.S.R}{k-1}\right) = \frac{h}{k-1}\sum_{i=1}^{k}\alpha_i^2 + \sigma_e^2.$

$E(S.S.C.) = E\{k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2\}$

$= E\{k\sum_{j=1}^{h}\bar{y}_{.j}^2 - hk\bar{y}_{..}^2\}$

$= k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) - hkE(\bar{y}_{..}^2)$

$= k\sum_{j=1}^{h}\left(\mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j\right) - hk\left(\mu^2 + \frac{\sigma_e^2}{hk}\right)$

$= hk\mu^2 + k\sum_{j=1}^{h}\beta_j^2 + h\sigma_e^2 + 2\mu\sum_{j=1}^{h}\beta_j - hk\mu^2 - \sigma_e^2$

$= k\sum_{j=1}^{h}\beta_j^2 + (h-1)\sigma_e^2.$ [since $\sum_{j=1}^{h}\beta_j = 0$].

Or $E(M.S.C.) = E\left(\frac{S.S.C}{h-1}\right) = \frac{k}{h-1}\sum_{j=1}^{h}\beta_j^2 + \sigma_e^2$

Now:

$E(S.S.E.) = E\{\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2\}$

$= E\{\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij}^2 + \bar{y}_{i.}^2 + \bar{y}_{.j}^2 + \bar{y}_{..}^2 - 2y_{ij}\bar{y}_{i.} - 2y_{ij}\bar{y}_{.j} + 2y_{ij}\bar{y}_{..} + 2\bar{y}_{i.}\bar{y}_{.j} - 2\bar{y}_{i.}\bar{y}_{..} -$

$2\bar{y}_{.j}\bar{y}_{..})\}$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) + h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) + k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2) -$

$2E\{\sum_{i=1}^{k}\bar{y}_{i.}\sum_{j=1}^{h}y_{ij}\} - 2E\{\sum_{j=1}^{h}\bar{y}_{.j}\sum_{i=1}^{k}y_{ij}\} + 2E\{\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{h}y_{ij}\} +$

$2E\{\sum_{i=1}^{k}\bar{y}_{i.}\sum_{j=1}^{h}\bar{y}_{.j}\} - 2E\{h\bar{y}_{..}\sum_{i=1}^{k}\bar{y}_{i.}\} - 2E\{k\bar{y}_{..}\sum_{j=1}^{h}\bar{y}_{.j}\}$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) + h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) + k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2) - 2h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) -$

$2k\sum_{j=1}^{h}E(\bar{y}_{.j}^2) + 2hkE(\bar{y}_{..}^2) + 2hkE(\bar{y}_{..}^2) - 2hkE(\bar{y}_{..}^2) - 2hkE(\bar{y}_{..}^2)$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}E(y_{ij}^2) - h\sum_{i=1}^{k}E(\bar{y}_{i.}^2) - k\sum_{j=1}^{h}E\left(\bar{y}_{.j}^2\right) + hkE(\bar{y}_{..}^2)$

$= \sum_{i=1}^{k}\sum_{j=1}^{h}\left(\mu^2 + \alpha_i^2 + \beta_j^2 + 2\mu\alpha_i + 2\mu\beta_j + 2\alpha_i\beta_j + \sigma_e^2\right) - h\sum_{i=1}^{k}(\mu^2 + \alpha_i^2 +$

$\frac{\sigma_e^2}{h} + 2\mu\alpha_i) - k\sum_{j=1}^{h}\left(\mu^2 + \beta_j^2 + \frac{\sigma_e^2}{k} + 2\mu\beta_j\right) + hk\left(\mu^2 + \frac{\sigma_e^2}{hk}\right)$

$= hk\mu^2 + h\sum_{i=1}^{k}\alpha_i^2 + k\sum_{j=1}^{h}\beta_j^2 + hk\sigma_e^2 - hk\mu^2 - h\sum_{i=1}^{k}\alpha_i^2 - k\sigma_e^2 - hk\mu^2 -$

$k\sum_{j=1}^{h}\beta_j^2 - h\sigma_e^2 + hk\mu^2 + \sigma_e^2$

[since $\sum_{i=1}^{k}\alpha_i = 0$ and $\sum_{j=1}^{h}\beta_j = 0$].

$= (hk - k - h + 1)\sigma_e^2 = (k-1)(h-1)\sigma_e^2$.

Or E(M.S.E.) = $E\left(\frac{S.S.E}{(k-1)(h-1)}\right) = \sigma_e^2$.

Thus, under H_{01}, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \Rightarrow \sum_{i=1}^{k}\alpha_i^2 = 0$.

Hence, E(M.S.R.) = σ_e^2 = E(M.S.E.).

Also, under H_{01}, S.S.R. follows a χ^2 distribution with k – 1 degrees of freedom and S.S.E. follows a χ^2 distribution with $(k-1)(h-1)$ degrees of freedom.

Hence, for testing H_{01}, the test statistic is given by $F_R = \frac{S.S.R/(k-1)}{S.S.E./(k-1)(h-1)} = \frac{M.S.R}{M.S.E}$, which will follow a central F distribution with k – 1 and $(k-1)(h-1)$ degrees of freedom.

Similarly, under H_{02}, $\beta_1 = \beta_2 = \cdots = \beta_h = 0 \Rightarrow \sum_{j=1}^{h}\beta_j^2 = 0$. Hence,

E(M.S.C.) = σ_e^2 = E(M.S.E.)

Also, under H_{02}, S.S.C. follows a χ^2 distribution with h – 1 degrees of freedom and S.S.E. follows a χ^2 distribution with $(k-1)(h-1)$ degrees of freedom.

Hence, for testing H_0, the test statistic is given by $F_C = \frac{S.S.C/(h-1)}{S.S.E./(k-1)(h-1)} = \frac{M.S.C}{M.S.E}$, which will follow a central F distribution with h – 1 and $(k-1)(h-1)$ degrees of freedom.

2.4.3 ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Rows	k – 1	S.S.R.= $h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$	M.S.R = $\frac{S.S.R.}{k-1}$	$F_R = \frac{M.S.R.}{M.S.E}$
Columns	h – 1	S.S.C.= $k\sum_{j=1}^{h}(\bar{y}_{.j} - \bar{y}_{..})^2$	M.S.C = $\frac{S.S.C.}{h-1}$	$F_C = \frac{M.S.C.}{M.S.E.}$
Error	(k –1)(h-1)	S.S.E.=$\sum_{i=1}^{k}\sum_{j=1}^{h}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$	M.S.E. = $\frac{S.S.E.}{(k-1)(h-1)}$	

Total	kh – 1	T.S.S. $=\sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2$	

- If $F_R < F_{\alpha/2,(k-1),(k-1)(h-1)}$, then H_{01} is accepted, hence we conclude that there is no significant difference between $\mu_{i's}$ otherwise H_{01} is rejected at level of significance α.
- If $F_{Cal,col} < F_{\alpha/2(h-1),(k-1)(h-1)}$ then H_{02} is accepted, hence we conclude that there is no significant difference between $\mu_{j's}$ otherwise H_{02} is rejected at level of significance α.

Critical Difference

If the mean effect due to factor A or factor B differ significantly, then we need to know about those pairs of means which differ significantly. For this we calculate the critical difference.

1. Critical Difference due to Row

   $CD_{row} = t_{\alpha/2(k-1),(h-1)}\sqrt{\frac{2MSE}{h}}$

   If $|\overline{y_{I.}} - \overline{y'_{I.}}| > CD_{row}$, then i^{th} and i'^{th} row means are said to differ significantly, otherwise not.
2. Critical Difference due to Column

   If $|\overline{y_{.J}} - \overline{y'_{.J}}| > CD_{col}$, then j^{th} and j'^{th} column means are said to differ significantly, otherwise not.

For Practical calculations

We have T.S.S $= \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\left(y_{ij}^2 + \bar{y}_{..}^2 - 2y_{ij}\bar{y}_{..}\right)$

$$= \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 + kh\,\bar{y}_{..}^2 - 2\bar{y}_{..}\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 + kh\,\bar{y}_{..}^2 - 2kh\bar{y}_{..}^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}^2 - kh\,\bar{y}_{..}^2$$

$$= \text{Raw Sum of Squares (RSS)} - kh\left(\frac{T_{..}}{hk}\right)^2$$

$$= RSS - \frac{T_{..}^2}{hk}$$

T.S.S = RSS $-$ Correction Factor (C.F.),

Where, C.F. = $\frac{T_{..}^2}{hk}$.

Similarly,

$$S.S.R = h\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = h\sum_{i=1}^{k}\bar{y}_{i.}^2 - hk\bar{y}_{..}^2$$

$$= h\sum_{i=1}^{k}\left(\frac{T_{i.}}{h}\right)^2 - C.F. = \frac{1}{h}\sum_{i=1}^{k}T_{i.}^2 - C.F.$$

$$S.S.C = k\sum_{j=1}^{h}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2 = k\sum_{j=1}^{h}\bar{y}_{.j}{}^2 - hk\bar{y}_{..}^2$$

$$= k\sum_{j=1}^{h}\left(\frac{T_{.j}}{k}\right)^2 - C.F = \frac{1}{k}\sum_{j=1}^{h}T_{.j}{}^2 - C.F\ .$$

S.S.E = T.S.S – S.S.R. – S.S.C.

Example: The following table gives monthly sales (in thousand rupees) of a certain firm in three states by its four salesmen. Set up analysis of variance table and test whether there is a significant difference between sales by the firm salesmen and sales in the three states.

States	*Salesmen*			
	I	*II*	*III*	*IV*
A	6	5	3	8
B	8	9	6	5
C	10	7	8	7

Solution: Let us take the hypothesis that there is no significant difference between the sales by the four salesmen, and there is no significant difference between sales in the three states.

States	*Salesmen*				*Total*
	I	*II*	*III*	*IV*	
A	6	5	3	8	22
B	8	9	6	5	28
C	10	7	8	7	32
Total	24	21	17	20	82

$\bar{X}$	8	7	5.67	6.67	

Correction Factor (CF) $= \frac{T^2}{N} = \frac{82^2}{12} = 560.333$

Total Sum of Squares (TSS) $= \sum_{i=1}^{3}\sum_{j=}^{4} y_{ij}^2 - CF = 602 - 560.333 = 41.667$

Sum of Squares between Columns (SSC) $= \frac{1}{3}\sum_{j=1}^{4} y_{.j}^2 - CF = \frac{1}{3}[24^2 + 21^2 + 17^2 + 20^2] - 560.333$

$= 568.67 - 560.333 = 8.337$

Sum of Squares between Rows (SSR) $= \frac{1}{4}\sum_{i=1}^{3} y_{i.}^2 - CF = \frac{1}{4}[22^2 + 28^2 + 32^2] - 560.333$

$= 573 - 560.333 = 12.667$

Sum of Squares due to Error(SSE) $= TSS - SSC - SSR = 41.667 - 8.337 - 12.667$

$= 20.663$

ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$F_{Cal.}$	$F_{Tab.}$
Rows	2	*12.667*	*6.334*	*1.839*	$F_{(2,6)} = 5.14$
Columns	3	*8.337*	*2.779*	0.807	$F_{(3,6)} = 4.76$
Error	*6*	*20.663*	*3.444*		
Total	11	*41.667*			

For the sales in the three states (rows), the calculated value of F is less than the tabulated value. Hence, there is no significant difference in the states as far as sales are concerned.

For the sales by the firm salesmen (columns), the calculated value of F is less than the tabulated value. Hence, we conclude that the sales of different salesmen do not differ significantly.

2.5 Self-Assessment Exercise

1. Explain the meaning of Analysis of Variance and state its basic assumptions.
2. For the one-way classified fixed effect model, $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$; (i=1,2,…,k; j=1,2,…,n) where the symbols having their usual meanings. Obtain:
   i. The estimates of the parameters $\mu + \alpha_i$

ii. The expectations of the various sum of squares.

iii. Give the ANOVA table

3. Give the fixed effect mathematical model for two-way classification with one observation per cell, stating clearly the assumptions involved. Also obtain:

   i. The estimates of the parameters in the model

   ii. The variance of the estimates

   iii. The expectation of the various sum of squares

   iv. ANOVA Table

4. Data collected on the effect of four fixed types of television tube coating on the conductivity of the tubes. Do an analysis of variance on these data and test the hypothesis that the four coatings yield the same average conductivity.

I	56	55	62	59	60
II	64	61	50	55	56
III	45	46	45	39	43
IV	42	39	45	43	41

5. A trucking company wishes to test the average life of each of the four brands of tyres. The company uses all brands on randomly selected trucks. The records showing the lives (thousands of miles) of tyres are as given in the table below. Test the hypothesis that the average life for each brand of tyres is the same.

Brand-1	20	23	18	17	
Brand-2	19	15	17	20	16
Brand-3	21	19	20	17	16
Brand-4	15	17	16	18	

6. Three different methods of analysis M_1, M_2 and M_3 are used to determine in parts per million the amount of a certain constituent in the sample. Each method is used by five analysts, and the results are given in the table below. Do the results indicate a significant variation either between the methods or between the analysts?

Analyst	*Method*		
	M_1	M_2	M_3
1	7.5	7.0	7.1
2	7.4	7.2	6.7
3	7.3	7.0	6.9
4	7.6	7.2	6.8
5	7.4	7.1	6.9

2.6 Summary

This unit makes imparts knowledge about the concept of analysis of variance (ANOVA) and teaches how to perform the analysis of variance in one-way and two-way classified data with equal (one) observation per cell.

2.7 References

- Gun, A.M., Gupta, M.K. and Dasgupta, B. (2013). *Fundamentals of Statistics Vol.1 & II*. World Press.
- Gupta, S.C. and Kapoor, V.K. (2008). *Fundamentals of Applied Statistics (4th ed.)*, Sultan Chand and Sons.
- Mukhopadhyay, P (2011). *Applied Statistics* (2nd edition revised reprint), Books & Allied (P) Ltd.

2.8 Further Reading

- Bapat, R. (2012). *Linear Algebra and Linear Models*. Springer.
- Hogg, R. V., McKean, J. W., and Craig, A. T. (2020). *Introduction to Mathematical Statistics*. Pearson.
- Mood, A. H., Bose, D. C., & Graybill, F. A. (1984). *Introduction to the theory of Statistics*. McGraw-Hill.

UNIT:3 ANALYSIS OF VARIANCE- II

Structure

3.1 Introduction

Experiments are performed to draw inferences about an entire population based on a few observations. If the experiments are perfectly repetable and the important factors giving rise to the results are perfectly separable, then the analysis and interpretation of results become relatively easy. However, experiments are often conducted so that the effect of one factor is (unknowingly) mixed up with the effect of a factor not considered in the experiment. These reasons, among others, make the analysis of the data from an experiment difficult. The role of statistics in experimental design is to separate the observed differences into those caused by various factors and those due to random fluctuations. The classical method used to separate these differences is *Analysis of Variance* or *ANOVA*.

When a set of observations is spread out across the different levels of two factors at the same time. Suppose that two factors A and B vary in an experiment, the factor A have p levels

A_1, A_2, …. A_p, and the factor B have q levels B_1, B_2, …. B_q. As an example, the factor A may be the variety of paddy (different varieties being A_1, A_2, …. A_p) and B may be the location (block) in the rural part of a distinct (different locations being B_1, B_2, …. B_q, where these varieties of crop are cultivated.

In such a two-factor experiments, the observations can be arranged in a two-way layout or a pxq table, where each row corresponds to a level Ai of A and each column to a level Bj, of B. Let n_{ij} be the number of observations in the cell (i,j) and y_{ijk} be the value of k^{th} observation on the $(i,j)^{th}$ cell, k = 1, 2,… n_{ij}; p=1,2,… p; j=1,2,… q. In the above example, y_{ijk} may be yield of paddy on the kth plot in the j^{th} location on which the i^{th} variety of paddy has been sown. We assume that the plots are of the same shape and size of unit area, and the k^{th} plot has been chosen randomly out of all such plots in the j^{th} location. If $n_{ij} \geq 1, \forall(i,j)$, the layout is called a *complete layout*. In an *incomplete layout*, n_{ij} = 0 for some (i,j).

ANOVA is performed in such a situation as a statistical method to find and measure the sources of variation. An extension of traditional two-way ANOVA called ANOVA in two-way classification with m-observations per cell is used when more than one observation is taken for each combination of factors in a two-way classification. This method works best when there are multiple readings to make the statistical analysis more reliable.

3.2 Objectives

After going through this unit, you should be able to:

- Perform the analysis of variance in a two-way classified data with m-observations per cell,
- Conduct Tukey's Test for Non-Additivity for Two-way layout with one observation per cell,
- Understand the concept of Analysis of Covariance (ANCOVA) for one-way and two-way classified data.

3.3 Analysis of Variance: Two-Way Classification with m-Observations Per Cell

In this case of two-way classified data with one observation per cell, we are not able to obtain an estimate of, or more make a test for the interaction effect. However, if some or all of the cells contain more than one observation, then we can estimate or test for the interaction effect. Here, we assume that there is an equal number, say m-observations in each cell. Let the m-observations be in the ij^th^ cell and denoted by $y_{ij1}, y_{ij2}, \ldots\ldots, y_{ijk} \ldots\ldots y_{ijm}$. Thus, y_{ijk} denotes the k^{th} observation for the i^{th} level of factor A and j^{th} level of factor B.

Factor A	***Factor B***				
	B_1	**……….**	***B_j***	**……….**	***B_q***
A_1	y_{111} y_{112} ……. y_{11m}		y_{1j1} ……. y_{1jm}		y_{1q1} ……. y_{1qm}
A_2	y_{211} y_{212} ……. y_{21m}		y_{2j1} ……. y_{2jm}		y_{2q1} ……. y_{2qm}
⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
A_i	y_{i11} y_{i12} ……. y_{i1m}		y_{ij1} ……. y_{ijm}		y_{iq1} ……. y_{iqm}
⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
A_p	y_{p11} y_{p12} ……. y_{p1m}		y_{pj1} ……. y_{pjm}		y_{pq1} ……. y_{pqm}

The above defined scheme is known as “Two Way Classification of data with m-observations per cell”.

3.3.1 Statistical Analysis

The total variation in the observation can be split into the following four components:

(i) The variation due to factor A

(ii) The variation due to factor B

(iii) The variation due to interaction effect AB

(iv) The variation due to random effect

Hence the mathematical model is given by:

$$y_{ijk} = \mu_{ijk} + e_{ijk}; \quad i = 1,2,\cdots,p; j = 1,2,\cdots,q; k = 1,2,\cdots,m \qquad (1)$$

Where,

μ_{ijk} be the true value for the ij^{th} cell and e_{ijk} be the error and $e_{ijk} \sim N(0, \sigma^2)$.

Now, μ_{ijk} can be decomposed as:

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Where,

$\sum_i \alpha_i = 0 = \sum_j \beta_j$

$\sum_i \gamma_{ij} = 0 \ \forall j \ and \ \sum_j \gamma_{ij} = 0 \ \forall i$

Now, our linear model (1) can be re-written as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (2)$$

Where,

μ is the general mean effect

α_i is the effect due to i^{th} level of factor A; $i = 1,2,\cdots,p$;

β_j is the effect due to j^{th} level of factor B; $i = 1,2,\cdots,q$;

γ_{ij} is the interaction effect between i^{th} level of factor A and j^{th} level of factor B, and

$e_{ij}'s$ are the error effect due to chance and these are assumed to be *iid*

Here, we want to test the equality (homogeneity) of the different level of factor A as well as factor B and independency of A and B. Thus, our hypotheses are:

H_{0A}: $\mu_{10} = \mu_{20} = \ldots\ldots = \mu_{p0} = \mu \ (=) \ \alpha_1 = \alpha_2 = \ldots\ldots = \alpha_p = 0$

H_{0B}: $\mu_{01} = \mu_{02} = \ldots\ldots = \mu_{0q} = \mu \ (=) \ \beta_1 = \beta_2 = \ldots\ldots = \beta_q = 0$

H_{AB}: $\gamma_{ij} = 0, \forall \ i \ \& \ j$

Against,

H_{1A}: At least two means are not same.

H_{1B}: At least two means are not equal.

H_{AB}: $\gamma_{ij} \neq 0, \forall \ i \ \& \ j$

3.3.2 Least Square Estimates

For testing above hypotheses, we need least square estimates of $\mu, \ \alpha_i, \ \beta_j \ and \ \gamma_{ij}$. Thus, the least square estimates can be obtained by minimizing the residual sum of square as:

$$S = \sum_i^p \sum_j^q \sum_k^m e_{ijk}^2 = \sum_i^p \sum_j^q \sum_k^m (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

The normal equations are:

$\frac{dS}{d\mu} = 0, \frac{dS}{d\alpha_i} = 0, \frac{dS}{d\beta_j} = 0, \frac{dS}{d\gamma_{ij}} = 0$

Now,

$\frac{dS}{d\mu} = 0 \Rightarrow \hat{\mu} = \bar{y}_{\ldots}$

$\frac{dS}{d\alpha_i} = 0 \Rightarrow \widehat{\alpha_i} = \bar{y}_{i..} - \bar{y}_{...}$

$\frac{dS}{d\beta_j} = 0 \Rightarrow \widehat{\beta_j} = \bar{y}_{.j.} - \bar{y}_{...}$

$\frac{dS}{d\gamma_{ij}} = 0 \Rightarrow 2\sum_k (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})(-1) = 0$

$\Rightarrow m\gamma_{ij} = y_{ij.} - m(\alpha_i + \beta_j) - m\mu$

$\Rightarrow \gamma_{ij} = \bar{y}_{ij.} - \mu - \alpha_i - \beta_j$

$\Rightarrow \widehat{\gamma_{ij}} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$

Now, substituting all these estimates in equation (2), we get:

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.})$$

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.})$$

Now, squaring both sides and summing over all observations, we get:

$$\sum_i^p \sum_j^q \sum_k^m (y_{ijk} - \bar{y}_{...})^2 = \sum_i^p \sum_j^q \sum_k^m \left[(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.})\right]^2$$

$$\sum_i^p \sum_j^q \sum_k^m (y_{ijk} - \bar{y}_{...})^2 = \sum_i^p \sum_j^q \sum_k^m (\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_i^p \sum_j^q \sum_k^m (\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_i^p \sum_j^q \sum_k^m (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_i^p \sum_j^q \sum_k^m (\bar{y}_{ijk} - \bar{y}_{ij.})^2$$

Since the product term will vanish, hence:

$$\sum_i^p \sum_j^q \sum_k^m (y_{ijk} - \bar{y}_{...})^2 = qm \sum_i^p (\bar{y}_{i..} - \bar{y}_{...})^2 + pm \sum_j^q (\bar{y}_{.j.} - \bar{y}_{...})^2 + m \sum_i^p \sum_i^q (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_i^p \sum_j^q \sum_k^m (\bar{y}_{ijk} - \bar{y}_{ij.})^2$$

TSS = SSA + SSB + SS(AB) + SSE

Total Sum of Square = Sum of Square due to Factor A + Sum of Square due to Factor B + Sum of Square due to Interaction between Factor A and Factor B + Sum of Square due to Error

Degrees of Freedom

TSS has (n-1) degree of freedom

SSA has (p-1) degree of freedom

SSB has (q-1) degree of freedom

SS(AB) has (p-1)(q-1) degree of freedom

SSE has pq(m-1) degree of freedom

In this case, we see that the degree of freedom is also additive in nature.

n – 1 = (p – 1) + (q- 1) + (p – 1)(q – 1) + pq(m -1)

Corresponding mean sum of squares are obtained as:

$MSA = \frac{SSA}{p-1}$ $\quad$ $MSB = \frac{SSB}{q-1}$

$MS(AB) = \frac{SS(AB)}{(p-1)(q-1)}$ $\quad$ $MSB = \frac{SSE}{pq(m-1)}$

F-test Statistic

Now to obtain appropriate test statistics to test the null hypothesis H_{OA}, H_{OB} and H_{OAB}, we find the expectation of mean sum of square from model (2).

We have by summing model (2) over j and k and dividing by mq:

$$\bar{y}_{i..} = \mu + \alpha_i + \bar{e}_{i..} \quad (3)$$

Summing model (2) over i and k and dividing by pm:

$$\bar{y}_{.j.} = \mu + \beta_j + \bar{e}_{.j.} \quad (4)$$

Similarly,

$$\bar{y}_{ij.} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \bar{e}_{ij.}$$

(5)

And

$$\bar{y}_{...} = \mu + \bar{e}_{...} \quad (6)$$

Then,

$$SSA = mq \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$= mq \sum_i (\mu - \alpha_i - \bar{e}_{i..} - \mu - \bar{e}_{...})^2$$

$$= mq \sum_i (\alpha_i + (\bar{e}_{i..} - \bar{e}_{...}))^2$$

$$= mq \sum_i \alpha_i^2 + mq \sum_i (\bar{e}_{i..} - \bar{e}_{...})^2 + 2mq \sum_i (\bar{e}_{i..} - \bar{e}_{...})\alpha_i$$

Now,

$$E(SSA) = mq \sum_i \alpha_i^2 + mq \sum_i E(\bar{e}_{i..} - \bar{e}_{...})^2 + 2mq \sum_i E(\bar{e}_{i..} - \bar{e}_{...})\alpha_i$$

$$= mq \sum_i \alpha_i^2 + mq \sum_i E\left[\bar{e}_{i..}^{\ 2} + \bar{e}_{...}^{\ 2} - 2\bar{e}_{i..} * \bar{e}_{...}\right] + 0$$

$$= mq\sum_i \alpha_i{}^2 + mq\left[\sum_i E(\bar{e}_{i..}{}^2) - pE(\bar{e}_{...}{}^2)\right]$$

$$= mq\sum_i \alpha_i{}^2 + mq\sum_i \frac{\sigma_e^2}{mq} - mpq * \frac{\sigma_e^2}{mpq}$$

$$= mq\sum_i \alpha_i{}^2 + \sigma_e^2 + p\sigma_e^2$$

Now,

SSB = $mp\sum_j \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$

$$= mp\sum_j \left(\mu + \beta_j + \bar{e}_{.j.} - \mu - \bar{e}_{...}\right)^2$$

$$= mp\sum_j \left[\beta_j + \left(\bar{e}_{.j.} - \bar{e}_{...}\right)\right]^2$$

$$= mp\sum_j \beta_j^2 + mp\sum_j \left(\bar{e}_{.j.} - \bar{e}_{...}\right)^2$$

Now,

E(SSB) $= mp\sum_j \beta_j^2 + mp\sum_j E\left(\bar{e}_{.j.} - \bar{e}_{...}\right)^2$

$$= mp\sum \beta_j^2 + mp\sum_j E\left[\bar{e}_{.j.}^2 + \bar{e}_{...}^2 - 2\ \bar{e}_{.j.}\,\bar{e}_{...}\right]$$

$$= mp\sum_j \beta_j^2 + mp\ E\left[\sum_j \bar{e}_{.j.}^2 + q\ \bar{e}_{...}^2 - 2\ q\ \bar{e}_{...}^2\right]$$

$$= mp\sum_j \beta_j^2 + mp\ \sum_j E\left[\bar{e}_{.j.}^2\right] - q\ mp\ E\left[\bar{e}_{...}^2\right]$$

$$= mp\sum_j \beta_j^2 + mp\frac{\sum_j \sigma_e^2}{mp} - pqm\frac{\sigma_e^2}{pqm}$$

$$= mp\sum_j \beta_j^2 + q\ \sigma_e^2 - \sigma_e^2$$

$$= (q-1)\sigma_e^2 + mp\ \sum_j \beta_j^2$$

Again,

SS(AB) $= m\sum_i\sum_j \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right)^2$

$$= m\sum_i\sum_j [(\mu + \alpha_i + \beta_j + \gamma_{ij} + \bar{e}_{ij.}) - (\mu + \alpha_i + \bar{e}_{i..}) - \left(\mu + \beta_j + \bar{e}_{.j.}\right) + (\mu + \bar{e}_{...})]^2$$

$$= m\sum_i\sum_j \left[\gamma_{ij} + \left(\bar{e}_{ij.} - e_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right)\right]^2$$

$= m\sum_i\sum_j \gamma_{ij}^2 + m\sum_i\sum_j\left(\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right)^2 + 2m\sum_i\sum_j \gamma_{ij}\left(\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right)^2$

Now,

E[MS(AB)] $= m\sum_i\sum_j \gamma_{ij}^2 + mE[\sum_i\sum_j\left(\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right)^2] + 2m\sum_i\sum_j \gamma_{ij}E\left[\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right]$

$= m\sum_i\sum_j \gamma_{ij}^2 + m\sum_i\sum_j E\left[\bar{e}_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}_{...}\right]^2 + 0$

$= m\sum_i\sum_j \gamma_{ij}^2 + m\,E[\sum_i\sum_j \bar{e}_{ij.}^2 + \sum_i\sum_j \bar{e}_{i..}^2 + \sum_i\sum_j \bar{e}_{.j.}^2 + \sum_i\sum_j \bar{e}_{...}^2 - 2\sum_i\sum_j \bar{e}_{ij.}\bar{e}_{i..} + 2\sum_i\sum_j \bar{e}_{ij.}\bar{e}_{.j.} + 2\sum_i\sum_j \bar{e}_{ij.}\bar{e}_{...} + 2\sum_i\sum_j \bar{e}_{i..}\bar{e}_{.j.} - 2\sum_i\sum_j \bar{e}_{i..}\bar{e}_{...} - 2\sum_i\sum_j \bar{e}_{.j.}\bar{e}_{...}]$

$= m\sum_i\sum_j \gamma_{ij}^2 + mE[\sum_i\sum_j \bar{e}_{ij.}^2 + q\sum_i \bar{e}_{i..}^2 + p\sum_j \bar{e}_{.j.}^2 + pq\,\bar{e}_{...}^2 - 2q\sum_i \bar{e}_{i..}^2 - 2p\sum_j \bar{e}_{.j.}^2 + 2pq\bar{e}_{...}^2 - 2pq\bar{e}_{...}]$

$= m\sum_i\sum_j \gamma_{ij}^2 + m\sum_i\sum_j E\left[\bar{e}_{ij.}^2\right] - mq\sum_i E[\bar{e}_{i..}^2] - mp\sum_j E\left[\bar{e}_{.j.}^2\right] + mpqE[\bar{e}_{...}^2]$

$= m\sum_i\sum_j \gamma_{ij}^2 + m\sum_i\sum_j \frac{\sigma_e^2}{m} - mq\sum_i \frac{\sigma_e^2}{mq} - mp\sum_j \frac{\sigma_e^2}{mp} + \frac{mpq\sigma_e^2}{mpq}$

$= m\sum_i\sum_j \gamma_{ij}^2 + pq\sigma_e^2 - p\sigma_e^2 - q\sigma_e^2 + \sigma_e^2$

$= (pq - p - q + 1)\sigma_e^2 + m\sum_i\sum_j \gamma_{ij}^2$

$= (p-1)(q-1)\sigma_e^2 + m\sum_i\sum_j \gamma_{ij}^2$

Now,

E [MS(AB)] $= E\left[\frac{SS(AB)}{(p-1)(q-1)}\right]$

$= \sigma_e^2 + m\sum_i\sum_j \gamma_{ij}^2$

Again,

SSE $= \sum_i\sum_j\sum_k\left(e_{ijk} - \bar{e}_{ij.}\right)^2$

$= \sum_i\sum_j\sum_k e_{ijk}^2 + \sum_i\sum_j\sum_k \bar{e}_{ij.}^2 - 2\sum_i\sum_j\sum_k e_{ijk}\bar{e}_{ij.}$

$= \sum_i \sum_j \sum_k e_{ijk}^2 + m \sum_i \sum_j \bar{e}_{ij.}^2 - 2m \sum_i \sum_j \bar{e}_{ij.}^2$

$= \sum_i \sum_j e_{ijk}^2 - m \sum_i \sum_j \bar{e}_{ij.}^2$

Now,

E[SSE] = $\sum_i \sum_j \sum_k E\left[e_{ijk}^2\right] - m \sum_i \sum_j E\left[\bar{e}_{ij.}^2\right]$

$= \sum_i \sum_j \sum_k \sigma_e^2 - m \sum_i \sum_j \frac{\sigma_e^2}{m}$

$= pqm\sigma_e^2 - pq\sigma_e^2$

$= pq(m-1)\sigma_e^2$

Mean Sum of Square

Dividing sum of squares by its degree of freedom, we get corresponding various mean sum of square,

E[MSA] $= E\left[\frac{SSA}{(p-1)}\right] = \sigma_e^2 + \frac{mq}{p-1} \sum_i \alpha_i^2$

$= \sigma_e^2 + \phi_1(\alpha_i)$

When H_{oA} is true, then E[MSA] = E[SSE]

E[SSA] = σ_e^2

E[MSB] = $E\left[\frac{SSB}{q-1}\right] = \frac{1}{q-1} E[SSB]$

$= \sigma_e^2 + \frac{mp}{q-1} \sum_j \beta_j^2$

$= \sigma_e^2 + \phi_2(\beta_j)$

When H_{oB} is true then: E[MSB] = E[MSE]

E[MS(AB)] = $E\left[\frac{SS(AB)}{(p-1)(q-1)}\right]$

$= \sigma_e^2 + \frac{m}{(p-1)(q-1)} \sum_i \sum_j (\gamma_{ij})^2$

$= \sigma_e^2 + \phi_3(\gamma_{ij})$

When H_{oAB} is true then: E[MS(AB)] = E[MSE]

Hence, when H_{oA}, H_{oB} *and* H_{oAB} are true. we have:

E[MSA] = E[MSB] = E[MS(AB)] =E[SSE] = σ

And the corresponding test statistics are:

$$F_A = \frac{MSA}{MSE} \sim F_{\{(p-1),pq(m-1)\}\left(\frac{\alpha}{2}\right)}$$

$$F_B = \frac{MSB}{MSE} \sim F_{\{(q-1),pq(m-1)\}\left(\frac{\alpha}{2}\right)}$$

$$F_{AB} = \frac{MS(AB)}{MSE} \sim F_{\{(p-1)(q-1),pq(m-1)\}\left(\frac{\alpha}{2}\right)}$$

3.3.3 ANOVA Table

Source of Variation	***Degrees of freedom***	***Sum of Square***	***Mean Sum of Square***	***Variance Ratio***
Factor A	p-1	$SSA = mq \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$	$MSA = \frac{SSA}{p-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	q-1	SSB = $mp \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$	$MSB = \frac{SSB}{q-1}$	$F_B = \frac{MSB}{MSE}$
Interaction between A&B	(p-1)(q-1)	SS(AB) = $m \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$MS(AB) = \frac{SS(AB)}{(p-1)(q-1)}$	$F_{AB} = \frac{MS(AB)}{MSE}$
Error	pq (m-1)	SSE = $\sum_i \sum_j \sum_k (y_{ijk} - -\bar{y}_{ij.})^2$	$MSE = \frac{SSE}{pq(m-1)}$	
Total	n-1	TSS = $\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$		

Example: A motor company wishes to check the influences of tyre type and shock absorber settings on the roadholding of one of its cars. Two types of tyres are selected from the tyre manufacturer who normally provides tyres for the company’s new vehicles. A shock absorber with three possible settings is chosen from a range of shock absorbers deemed to be suitable for the car. An experiment is conducted by conducting roadholding tests using each tyre type and shock absorber setting. The (coded) data resulting from the experiment are given below.

Factor	***Shock Absorber Setting***		
Tyre	*B1=Comfort*	*B2=Normal*	*B3=Sport*
Type A1	5 6 8	8 5 3	6 9 12
Type A2	9 7 7	10 9 8	12 10 9

Derive the appropriate ANOVA table. State clearly any conclusions that may be drawn at the 5% level of significance.

Solution:

Factor	***Shock Absorber Setting***			***Total***
Tyre	*B1*	*B2*	*B3*	
Type A1	5 6 8	8 5 3	6 9 12	
	$y_{11.} = 19$	$y_{12.} = 16$	$y_{13.} = 27$	$y_{1..} = 62$
Type A2	9 7 7	10 9 8	12 10 9	
	$y_{21.} = 23$	$y_{22.} = 27$	$y_{23.} = 31$	$y_{2..} = 81$
Total	$y_{.1.} = 42$	$y_{.2.} = 43$	$y_{.3.} = 58$	$y_{...} = 143$

Total Sum of Squares (TSS) $= \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3} y_{ijk}^2 - \frac{y_{...}^2}{N} = 1233 - \frac{143^2}{18} = 96.944$

Sum of Squares due to Factor A (SSA) $= \sum_{i=1}^{2} \frac{y_{i..}^2}{qm} - \frac{y_{...}^2}{N} = \frac{62^2+81^2}{3*3} - \frac{143^2}{18}$

$$= \frac{10405}{9} - \frac{143^2}{18} = 20.056$$

Sum of Squares due to Factor B (SSB) $= \sum_{j=1}^{3} \frac{y_{.j.}^2}{pm} - \frac{y_{...}^2}{N} = \frac{42^2+43^2+58^2}{2*3} - \frac{143^2}{18}$

$$= \frac{6977}{6} - \frac{143^2}{18} = 26.778$$

Sum of Squares due to Interaction (SSAB) $= \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{y_{ij.}^2}{m} - \frac{y_{...}^2}{N} - SSA - SSB$

$$= \frac{19^2+16^2+27^2+23^2+27^2+31^2}{3} - \frac{143^2}{18} - 20.056 - 26.778 = 5.444$$

ANOVA Table

Source of Variation	***Degrees of freedom***	***Sum of Square***	***Mean Sum of Square***	***Variance Ratio***	
				$F_{Cal.}$	***$F_{Tab.}$***
Factor A	1	20.056	20.056	5.39	$F_{(1,12)} = 4.75$
Factor B	2	26.778	13.389	3.60	$F_{(2,12)} = 3.89$
Interaction AB	2	5.444	2.722	0.731	$F_{(2,12)} = 3.89$
Error	12	44.666	3.722		
Total	17	96.944			

The following conclusions may be drawn:

Interaction: There is insufficient evidence to support the hypothesis that interaction takes place between the factors.

Factor A: Since 5.39 > 4.75 we have sufficient evidence to reject the hypothesis that tyre type does not affect the roadholding of the car.

Factor B: Since 3.60 < 3.89 we do not have sufficient evidence to reject the hypothesis that shock absorber settings do not affect the roadholding of the car.

3.4	**Tukey's Test for Non-Additivity for Two-Way Layout with One Observation Per Cell**

The linear model for a two-way layout with one observation per cell is:

Rows	***Columns***						***Row Totals***	***Row Means***
	1	***2***	$\cdots$	***j***	$\cdots$	***h***		
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1h}	$T_{1.} = \sum_{j=1}^{h} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{h}$
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2h}	$T_{2.} = \sum_{j=1}^{h} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{h}$
$\vdots$ $\vdots$			$\vdots$ $\vdots$				$\vdots$ $\vdots$	$\vdots$ $\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ih}	$T_{i.} = \sum_{j=1}^{h} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{h}$
$\vdots$ $\vdots$			$\vdots$ $\vdots$				$\vdots$ $\vdots$	$\vdots$ $\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kh}	$T_{k.} = \sum_{j=1}^{h} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{h}$
Column Totals	$T_{.1} = \sum_{i=1}^{k} y_{i1}$	$T_{.2} = \sum_{i=1}^{k} y_{i2}$	$\cdots$	$T_{.j} = \sum_{i=1}^{k} y_{ij}$	$\cdots$	$T_{.h} = \sum_{i=1}^{k} y_{ih}$	$T_{..} = \sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}$	
Column Means	$\bar{y}_{.1} = \frac{T_{.1}}{k}$	$\bar{y}_{.2} = \frac{T_{.2}}{k}$	$\cdots$	$\bar{y}_{.j} = \frac{T_{.j}}{k}$	$\cdots$	$\bar{y}_{.h} = \frac{T_{.h}}{k}$		$\bar{y}_{..} = \frac{T_{..}}{hk}$

Let y_{ij} denote the value of the observation in the $(i, j)^{\text{th}}$ cell and suppose that y_{ij}'s are *iid* random variables, distributed according to $N(\mu_{ij}, \sigma_e^2)$. Then the mathematical model is:

$$y_{ij} = \mu_{ij} + e_{ij}; i = 1,2,\cdots,k; j = 1,2,\cdots,h$$

where $e_{ij}'s$ are the error effect due to chance and these are assumed to be *iid* random variables each following $N(0, \sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h.$

μ_{ij} is further split into:

(1) $\mu = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij} = \frac{1}{hk}\sum_{i=1}^{k}\sum_{j=1}^{h}\mu_{ij}$, the over all mean;

(2) the row effect $\alpha_i = \mu_{i.} - \mu$, where $\mu_{i.} = \frac{1}{h}\sum_{j=1}^{h}\mu_{ij}$; and

(3) the column effect $\beta_j = \mu_{.j} - \mu$, where $\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}$.

(4) the interaction effect γ_{ij} when the i^{th} level of first factor and j^{th} level of second factor occur simultaneously and is given by: $\gamma_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$, where $\sum_j \gamma_{ij} = 0\ \forall i = 1,2,\ldots.k$ and $\sum_i \gamma_{ij} = 0\ \forall j = 1,2,\ldots.h$

Obviously,

$\sum_{i=1}^{k}\alpha_i = \sum_{i=1}^{k}(\mu_{i.} - \mu) = \sum_{i=1}^{k}\mu_{i.} - k\,\mu = k\,\mu - k\,\mu = 0.$

Similarly,

$\sum_{j=1}^{h}\beta_j = \sum_{j=1}^{h}(\mu_{.j} - \mu) = \sum_{j=1}^{h}\mu_{.j} - h\,\mu = h\,\mu - h\,\mu = 0.$

Thus,

$\mu_{ij} = \mu + \mu_{i.} - \mu + \mu_{.j} - \mu + \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu = \mu + \alpha_i + \beta_j + \gamma_{ij}$

Hence the mathematical model is given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij};\ i = 1,2,\cdots,k;\ j = 1,2,\cdots,h \qquad (1)$$

Where,

y_{ij} is the observation of i^{th} row and j^{th} column; $i = 1,2,\cdots,k, j = 1,2,\cdots,h,$

μ is the general mean effect

α_i is the additive effect due to i^{th}row; $i = 1,2,\cdots,k;$

β_j is the additive effect due to j^{th}column; $i = 1,2,\cdots,k;$

γ_{ij} is the interaction effect when the i^{th} level of first factor and j^{th} level of second factor occur simultaneously, and

$e_{ij}'s$ are the error effect due to chance and these are assumed to be *iid* random variables each following $N(0,\sigma_e^2)$; $i = 1,2,\cdots,k, j = 1,2,\cdots,h.$

The side conditions are $\sum_{i=1}^{k}\alpha_i = \sum_{j=1}^{h}\beta_j = 0.$

The null hypotheses to be tested in this case are:

$H_{01}: \alpha_1 = \alpha_2 = \ldots\ldots \alpha_k = 0; H_{02}: \beta_1 = \beta_2 = \ldots\ldots \beta_k = 0; H_{03}: \gamma_{ij} = 0$;i=1,2,…k; j=1,2,,…k

First of all, we have to test H_{o3}, since we cannot test H_{o1} or H_{o2} (i.e., the hypotheses on the main effects), unless it has been established that the interaction effect is zero.

To test H_{o3}: $\gamma_{ij} = 0 \ \forall I \ and \ j$, Tukey developed a procedure known as Tukey's test. The technique consists in partitioning the error sum of squares (S_E^2) with (k-1)(h-1) d.f. into two components as follows:

i. $S_{E_1}^2$ $(= SSN)$ i.e., the sum of squares due to non-additivity (i.e., interaction) which has single d.f. and

ii. $S_{E_2}^2$ i.e., the balance error sum of squares which as (k-1)(h-1)-1 d.f. We have:

$$S_{E_1}^2 = S.S. due\ to\ non - additivity\ (SSN)$$

$$= \frac{[\sum_{i=1}^{k}\sum_{j=1}^{h} y_{ij}\bar{y}_{i.}\bar{y}_{.j} - y_{..}(S_A^2 + S_B^2 + \frac{\bar{y}_{..}^2}{kh})]}{khS_A^2.S_B^2} \text{ with 1 d.f.}$$

$$And, S_{E_2}^2 = S_E^2 - S_{E_1}^2 = S_E^2 - SSN, with\ (k-1)(h-1)\ d.f.$$

Test statistics for testing H_{o3}: $\gamma_{ij} = 0 \ \forall I \ and \ j$ is:

$$F = \frac{S_{E_1}^2/1}{S_{E_2}^2/(k-1)(h-1)} = \frac{S_{E_1}^2}{S_{E_2}^2} \sim F_{1,(k-1)(h-1)-1}$$

If $F > F_{1,(k-1)(h-1)-1}(\alpha)$, we reject H_{o3} at α level of significance otherwise we may accept H_{o3}.

3.5 Analysis of Covariance (ANCOVA)

The basic objective of the designs considered so far is to make the treatment comparisons with the greatest precision by reducing the experimental error through the powerful tool of local control. Analysis of Covariance (ANCOVA), like Randomized Block Design or Latin Square Design, is a technique of increasing the precision of design by reducing the experimental error.

ANCOVA is a technique in which it is possible to control certain sources of variation by taking additional observations on each of the experimental units. Let us suppose that in an experiment, y is the response variable and x is another variable which is linearly related to y. Moreover, x cannot be controlled by the experimenter but can be observed along with the y's. The variable x is called the *covariate / concomitant / independent / ancillary* variable. In ANCOVA, we adjust for the variation in the response variable (y) for the linear regression (effect) of the independent variable (x). If this is not done, then the error mean square will be inflated due to the linear effect of x, thus making it difficult to detect the true differences in the response variable. ANCOVA procedure is a combination of the Analysis of Variance (ANOVA) and the regression analysis. Whenever it is possible to take additional observations on one or more the variables from each of the experimental units in the design along with the response variable under study, the ANCOVA technique has proved to be useful in many fields of research.

Some examples of ANCOVA are:

- Suppose we want to compare the effect of some rations (diets) on the weight of animals. We can analyze the data by performing the ANCOVA by regarding the final weight of the animals taking the ration, after a specified period as the response variable (y) and the initial weight of the animals at the time of starting the experiment as the concomitant variable (x). To ensure that the real differences in the final weights (y) are due to rations, we must adjust for the linear effect of the initial weight (x) on y.
- Suppose we want to compare the differences in the strength of the filament fibre (y) produced by different machines. Obviously, y depends on the thickness (x) of the fibre-thicker the fiber, stronger it is. The effect of the thickness (x) on the strength (y) can be eliminated by performing ANCOVA between the response variable (y) and the concomitant variable (x) for testing the differences in the strength of the fibre produced by different machines.
- In plant breeding experiments, suppose an equal number of seeds are sown per plot but at the time of harvest, the final number of plants in each plot will not be same due to certain reasons like non-germination of certain seeds, early death of certain plants, attack by birds/cattle etc.) and will vary from plot to plot. The yield (y) of a crop from different plots may depend on the number of plants (x) per plot. To study the real differences between the yields, we adjust for the linear effect of the number of plants per plot by performing ANCOVA by regarding the yield per plot (y) as the response variable and the number of plants per plot (x) as the concomitant variable.

Note: The concomitant variable need not necessarily be measurable. Even if it is a quality characteristic which cannot be measured quantitatively., intelligence, poverty, indifference, good/bad, presence/absence etc., but can be suitably converted into numerical scores, the use of ANCOVA results in a considerable increase in precision.

3.4.1 ANCOVA for One-Way Classification with a Single Concomitant Variable in C.R.D. Layout

Let us suppose that we are comparing v treatments $t_1, t_2 \ldots \ldots t_i, i^{(th)}$ treatment replicated $r_i, (i = 1,2 \ldots \ldots \ldots . p)$ times so that $n = \sum_{i=1}^{v} r_i$,is the total number of experimental units. Further suppose that the experiment is conducted with a CRD layout.

Suppose that along with the response (dependent) variable y we consider a single concomitant variable x. then the linear ANOCOVA model will consist of the sum of two components-one is the same component as in ANOVA and the second component is due to the regression of y on the concomitant variable x.

Then assuming a linear relationship between the response variable(y) and the concomitant variable (x), the appropriate statistical model (for fixed effects) for ANOCOVA for CRD with one concomitant variable is given by:

$$y_{(ij)} = \mu + \alpha_i + \beta\left(x_{(ij)} - \bar{x}..\right) + e_{(ij)} \qquad (1)$$

Where,

μ is the general mean effect

α_i is the (fixed) additional effect due to i^{th} treatment (i=1,2,….p)

β is the regression coefficient of y on x

x_{ij} be the concomitant variable corresponding to the response variable y_{ij} and

e_{ij} be the error and $e_{ij} \sim N(0, \sigma_e^2)$

Obviously $\sum \alpha_i = 0$

Estimation of Parameters

Here the residual sum of squares is given by

$$SSE = \sum_i \sum_j e_{(ij)}^2 = \sum_i \sum_j (y_{(ij)} - \mu - \alpha_i - \beta\left(x_{(ij)} - \bar{x}..\right))^2 \qquad (2)$$

To estimate we need

$$\frac{\partial(SSE)}{\partial\mu} = 0, \frac{\partial(SSE)}{\partial\alpha_i} = 0, \frac{\partial(SSE)}{\partial\beta} = 0$$

On solving this equation, we get

$$\hat{\mu} = \bar{y}.. \qquad (3)$$

$$\hat{\alpha_i} = (\bar{y}. - \bar{y}..) - \hat{\beta}(\bar{x}_i. - \bar{x}..) \qquad (4)$$

Also

$$\frac{\partial(SSE)}{\partial \beta} = 0$$

$$\sum_i \sum_j (y_{(ij)} - \bar{y}..) - (\overline{y_i}. - \bar{y}..) + \hat{\beta}((\bar{x}_i. - \bar{x}..) - \hat{\beta}(x_{(ij)} - \bar{x}..)](x_{(ij)} - \bar{x}..) = 0$$

$$\sum_i \sum_j [y_{(ij)} - \overline{y_i} \;\; - \hat{\beta}(x_{(ij)} - \bar{x}_i.)](x_{(ij)} - \bar{x}..) = 0$$

$$\sum_i \sum_j [y_{(ij)} - \overline{y_i.} \; - \hat{\beta}(x_{(ij)} - \bar{x}_i.)][(x_{(ij)} - \bar{x}_i.) + (\bar{x}_i. - \bar{x}..)] = 0$$

$$\sum_i \sum_j [(y_{(ij)} - \overline{y_i.})(x_{(ij)} - \bar{x}_i.)] \; + \sum_i \sum_j [(y_{(ij)} - \overline{y_i.})(\bar{x}_i. - \bar{x}..) - \hat{\beta}\left[\sum_i \sum_j \left(x_{ij} - \; \bar{x}_i.\right)^2 - \sum_i \sum_j [(x_{(ij)} - \overline{x_i.})(\bar{x}_i. - \bar{x}..)\right] = 0$$

$$\sum_i \sum_j [(y_{(ij)} - \overline{y_i.})(x_{(ij)} - \bar{x}_i.)] \; - \hat{\beta} \sum_i \sum_j \left(x_{ij} - \;\; \bar{x}_i.\right)^2 = 0$$

$$\hat{\beta} = \frac{\sum_i \sum_j [(y_{(ij)} - \overline{y_i.})(x_{(ij)} - \bar{x}_i.)]}{\sum_i \sum_j \left(x_{ij} - \; \bar{x}_i.\right)^2}$$

Now let

$$E_{(xx)} = \sum_i \sum_j \left(x_{ij} - \;\; \bar{x}_i.\right)^2$$

$$E_{(yy)} = \sum_i \sum_j \left(y_{ij} - \;\; \bar{y}_i.\right)^2$$

$$E_{(xy)} = \sum_i \sum_j [(y_{(ij)} - \overline{y_i.})(x_{(ij)} - \overline{x_i.})$$

$$E_{(xy)} = \sum_i \sum_j (x_{(ij)} - \overline{x_i.})\,(y_{(ij)} - \overline{y_i.})$$

$$T_{(xx)} = \sum \mathrm{n_i}\,(\overline{x_i.} - \bar{x}..)^2$$

$$T_{(yy)} = \sum \mathrm{n_i}\,(\overline{y_i.} - \bar{y}..)^2$$

$$T_{(xy)} = \sum \mathrm{n_i}\,(\overline{y_i.} - \bar{y}..)(\overline{x_i.} - \bar{x}..)$$

$$T_{(xy)} = \sum \mathrm{n_i}\,(\overline{x_i.} - \bar{x}..)(\overline{y_i.} - \bar{y}..)$$

$$E'_{(xx)} = \sum_i \sum_j \left(x_{ij} - \bar{x}_{..}\right)^2$$

$$E'_{(yy)} = \sum_i \sum_j \left(y_{ij} - \bar{y}_{..}\right)^2$$

$$E'_{(xy)} = \sum_i \sum_j [(y_{(ij)} - \bar{y}..)(x_{(ij)} - \bar{x}..)$$

$$E'_{(xy)} = \sum_i \sum_j (x_{(ij)} - \bar{x}..)\,(y_{(ij)} - \bar{y}..)$$

Therefore,

$$\hat{\beta} = \frac{E_{(xy)}}{E_{(xx)}}$$

The value of α_i,μ and β on putting in model 2) we get the unrestricted residual sum of square obtained for the above model is:

$$\text{SSE} = \sum_i \sum_j [(y_{(ij)} - \bar{y}..) - (\bar{y_i}. - \bar{y}..) + \hat{\beta}((\bar{x}_i. - \bar{x}..) - \hat{\beta}\left(x_{(ij)} - \bar{x}..\right)]^2 = 0$$

$$\text{SSE} = \sum_i \sum_j [(y_{(ij)} - \bar{y_i}. - \hat{\beta}\left(x_{(ij)} - \bar{x}_i.\right)]^2 = 0$$

$$\text{SSE} = \sum_i \sum_j \left(y_{ij} - \bar{y}_{i.}\right)^2 + \widehat{\beta^2}\,\left(x_{(ij)} - \bar{x}_i.\right)]^2 - 2\beta \sum_i \sum_j (x_{(ij)} - \overline{x_i.})\,(y_{(ij)} - \overline{y_i.}) = 0$$

$$E_{(yy)} + \left(\frac{(E_{(xy)})}{E_{(xx)}}\right)^2 * E_{(xx)} - \left(\frac{2 * E_{(xy)}}{E_{(xx)}}\right) * E_{(xy)}$$

$$SSE = E_{(yy)} - \left(\frac{E^2_{(xy)}}{E_{(xx)}}\right) \qquad (5)$$

SSE has $n - p - 1$ degrees of freedom i.e., $(n - 1) - (p - 1) - 1 = n - p - 1$

Here the null hypothesis H_0 is such that all the effects due to different treatments in the presence of concomitant variables are same i.e.

$$H_0\!: \alpha_1 = \alpha_2 = \cdots \ldots. = \alpha_p = 0$$

Under H_0 the model (1) reduces to:

$$\mathrm{y_{(ij)}} = \mu + \beta'\left(x_{(ij)} - \bar{x}..\right) + \mathrm{e_{(ij)}} \qquad (6)$$

And the error sum of square under H_0 is given by:

$$\text{SSE} *= \sum_i \sum_j e^2_{(ij)} = \sum_i \sum_j [(y_{(ij)} - \mu - \beta'\left(x_{(ij)} - \bar{x}..\right)]^2$$

For find the estimate μ and β′ we need:

$\frac{\partial(SSE*)}{\partial \mu} = 0$, $\frac{\partial(SSE*)}{\partial \beta'} = 0$

$\frac{\partial(SSE*)}{\partial \mu} = 0$

$\hat{\mu} = \bar{y}_{..}$

and

$\frac{\partial(SSE*)}{\partial \beta'} = \sum_i \sum_j [(y_{(ij)} - \mu - \beta'(x_{(ij)} - \bar{x}_{..})] (x_{(ij)} - \bar{x}_{..})=0$

$\sum_i \sum_j [(y_{(ij)} - \bar{y}_{..}) - \beta'(x_{(ij)} - \bar{x}_{..}) (x_{(ij)} - \bar{x}_{..})=0$

$\sum_i \sum_j [(y_{(ij)} - \bar{y}_{..})(x_{(ij)} - \bar{x}_{..}) - \beta'(x_{(ij)} - \bar{x}_{..})^2=0$

$\hat{\beta}' = \sum_i \sum_j [(y_{(ij)} - \bar{y}_{..})(x_{(ij)} - \bar{x}_{..})/ \sum_i \sum_j (x_{(ij)} - \overline{x_{i.}})^2$

$\hat{\beta}' = E'_{(xy)}/E'_{(xx)}$

Same as doing the previous procedure we can find SSE* i.e., the restricted residual sum of square (i.e., residual sum of square under H_0) is:

$SSE *= \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 - \hat{\beta}' \sum_i \sum_j (x_{(ij)} - \bar{x}_{..}) (y_{(ij)} - \bar{y}_{..})$

$= E_{(y'y)} - \frac{E'^2_{(xy)}}{E'_{(xx)}}$

Degree of freedom for $SSE*$ is $n - 1 - 1 = n - 2$

Thus, the sum of square due to treatment is:

$$SST = SSE^* - SSE$$

Degrees of freedom for SST = degree of freedom for SSE^* - degree of freedom for SSE

$$= n - 2 - (n - p - 1) = p - 1$$

$MST = (SST)/(p-1)$, $MSE = (SSE)/(n-p-1)$,

The appropriate test for testing H_0 is based on the test statistic F is given as

$$F = MST/MSE \sim F^{(\alpha)}_{p-1,(n-p-1)}$$

And H_0 is rejected at level of $\alpha * 100\%$ if $F^{(\alpha)}_{p-1,(n-p-1)}$, otherwise H_0 is accepted

ANOVA Table for One-Way Classification (CRD Layout)

Sources of variance	***Degree of freedom***	***Sum of Square***			***Estimate of β***	***Adjusted*** $SS_{(yy)}$	***Adjusted Degree of freedom***
		$SS_{(xx)}$	$SP_{(xy)}$	$SS_{(yy)}$			
Treatment	$p-1$	$T_{(xx)}$	$T_{(xy)}$	$T_{(yy)}$			
Error	$n-p$	$E_{(xx)}$	$E_{(xy)}$	$E_{(yy)}$	$\frac{E_{(xy)}}{E_{(xx)}}$	SSE	$n-p-1$
Total	$n-1$	$E'_{(xx)}$	$E'_{(xy)}$	$E'_{(yy)}$	$\frac{E'_{(xy)}}{E'_{(xx)}}$	SSE*	$n-2$
Difference (*Total – Error*)						SSE*- SSE	$p-1$

3.4.2 ANCOVA for Two-Way Classification with a Single Concomitant Variable in R.B.D. Layout

Suppose that we are comparing p treatments and each treatments replicated q times, so that n= pq be the total number of experimental units. Further suppose that the experiment is performed with a RBD layout. Here linear model will be

$$y_{ij} = \mu + \alpha_i + \theta_j + \beta(x_{ij} - \bar{x}_{..}) + e_{ij} \qquad (1)$$

Where,

μ is a general mean effect

α_i is additional effect due to i^{th} treatment over general mean effect

θ_j is additional effect due to j^{th} block over general mean effect

β is the regression coefficient of y on x

x_{ij} be the concomitant variable corresponding to the response variability

e_{ij} be the error and obviously.

$\sum_i \alpha_i = \sum_i \theta_j = 0$

Estimation of Parameters

Here the residual sum of square is:

$$SSE = \sum_i \sum_j e^2_{(ij)} = \sum_i \sum_j (y_{ij} - \mu - \alpha_i - \theta_j - \beta(x_{ij} - \bar{x}..)^2 \qquad (2)$$

To estimate μ, α_i, θ_j *and* β we need:

$\frac{\partial(SSE)}{\partial\mu} = 0, \frac{\partial(SSE)}{\partial\alpha_i} = 0, \frac{\partial(SSE)}{\partial\theta_j}\ \frac{\partial(SSE)}{\partial\beta} = 0$

Now, $\frac{\partial(SSE)}{\partial\mu} = 2\sum_i\sum_j[y_{ij} - \mu - \alpha_i - \theta_j - \beta(x_{ij} - \bar{x}..)](-1) = 0$

$\Rightarrow \sum_i\sum_j y_{ij} - \sum_i\sum_j \mu - \sum_i\sum_j \alpha_i - \sum_i\sum_j \theta_j - \beta\sum_i\sum_j (x_{ij} - \bar{x}..) = 0$

$\Rightarrow \sum_i\sum_j y_{ij} - pq\mu - 0 - 0 - 0 = 0$

$\Rightarrow \mu = \frac{1}{pq}\sum_i\sum_j y_{ij}$

$\Rightarrow \hat{\mu} = \bar{y}..$

Again

$\frac{\partial(SSE)}{\partial\alpha_i} = 2\sum_i\sum_j[y_{ij} - \mu - \alpha_i - \theta_j - \beta(x_{ij} - \bar{x}..)] = 0$

$\Rightarrow \sum_j y_{ij} - \sum_j \mu - \sum_j \alpha_i - \sum_j \theta_j - \beta\sum_j(x_{ij} - \bar{x}..) = 0$

$\Rightarrow \sum_j y_{ij} - q\hat{\mu} - q\alpha - 0 - q\beta(x_{i.} - \bar{x}..) = 0$

$\Rightarrow \overline{y_{i.}} - \hat{\mu} - \beta(\overline{x_{i.}} - \bar{x}..) = \alpha$

$\Rightarrow \hat{\alpha_i} = (\overline{y_{i.}} - \overline{y_{..}}) - \hat{\beta}(x_{i.} - \bar{x}..)$

Similarly

$\hat{\theta_j} = \left(\overline{y_{.j}} - \overline{y_{..}}\right) - \hat{\beta}\left(x_{.j} - \bar{x}..\right)$

Again

$\frac{\partial(SSE)}{\partial\beta} = 2\sum_i\sum_j[y_{ij} - \hat{\mu} - \hat{\alpha_i} - \theta_j - \hat{\beta}(x_{ij} - \bar{x}..)](x_{ij} - \bar{x}..) = 0$

$\Rightarrow \sum_i\sum_j[(y_{(ij)} - \bar{y}..) - (\overline{y_{i.}} - \overline{y_{..}}) + \hat{\beta}(x_{i.} - \bar{x}..) - (\overline{y_{.j}} - \overline{y_{..}}) + \hat{\beta}(x_{.j} - \bar{x}..) -$

$\hat{\beta}([(x_{(ij)} - \bar{x}..)](x_{ij} - \bar{x}..) = 0$

$\Rightarrow \sum_i\sum_j[(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)((x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..) - \beta\sum_i\sum_j(x_{(ij)}\ \overline{x_{i.}} - \overline{x_{.j}} +$

$\bar{x}..)^2 - \sum_i\sum_j[(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)\{(x_{i.} - \bar{x}..)(x_{.j} - \bar{x}..)\} - \beta\sum_i\sum_j(x_{(ij)}\ \overline{x_{i.}} -$

$\overline{x_{.j}} + \bar{x}..)\{(x_{i.} - \bar{x}..)(x_{.j} - \bar{x}..)\} = 0$

$$\Rightarrow \sum_i \sum_j [(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)((x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..) - \beta \sum_i \sum_j (x_{(ij)}\ \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)^2 - \sum_i \sum_j [(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)\{(x_{i.} - \bar{x}..) + \sum_i \sum_j [(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)(x_{.j} - \bar{x}..)\} - \beta \sum_i \sum_j (x_{(ij)}\ \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)\{(x_{i.} - \bar{x}..) + \beta \sum_i \sum_j (x_{(ij)}\ \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)(x_{.j} - \bar{x}..) = 0$$

$$\Rightarrow \sum_i \sum_j [(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)((x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..) - \beta \sum_i \sum_j (x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)^2 = 0$$

Other term will be vanishing because the sum of deviation about mean is zero.

$$\hat{\beta} = \frac{\sum_i \sum_j [(y_{(ij)} - \overline{y_{i.}} - \overline{y_{.j}} + \bar{y}..)((x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)}{\sum_i \sum_j (x_{(ij)} - \overline{x_{i.}} - \overline{x_{.j}} + \bar{x}..)^2}$$

Let

$$E_{xx} = \sum\sum\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)^2$$

$$E_{yy} = \sum\sum\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$$

$$E_{xy} = \sum\sum(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

$$\sum\sum\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})$$

$$T_{xx} = p\sum_j\left(\bar{x}_{.j} - \bar{x}_{..}\right)^2 = q\sum_i(\bar{x}_{i.} - \bar{x}_{..})^2$$

$$T_{yy} = p\sum_j\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$$

$$T_{xy} = p\sum_j(\bar{y}_{.j} - \bar{y}_{..})(\bar{x}_{i.} - \bar{x}_{..})$$

$$= p\sum_j(\bar{x}_{.j} - \bar{x}_{..})(\bar{y}_{.j} - \bar{y}_{..})$$

$$E'_{xx} = \sum\sum\left(x_{ij} - \bar{x}_{.j}\right)^2$$

$$E'_{yy} = \sum\sum\left(y_{ij} - \bar{y}_{.j}\right)^2$$

$$E'_{xy} = \sum\sum(x_{ij} - \bar{x}_{.j})(y_{ij} - \bar{y}_{.j})$$

Therefore,

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}}$$

Substituting the values of $\hat{\mu}, \hat{\alpha}_i, \hat{\theta}_j$ and $\hat{\beta}$ in equation (2) we get,

$$SSE = \sum\sum\left[y_{ij} - \bar{y}_{..} - (\bar{y}_{i.} - \bar{y}_{..}) + \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) - \left(\bar{y}_{.j} - \bar{y}_{..}\right) + \hat{\beta}\left(\bar{x}_{.j} - \bar{x}_{..}\right) - \hat{\beta}\left(x_{ij} - \bar{x}_{..}\right)\right]^2$$

$$= \sum\sum\left[\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right) - \hat{\beta}\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)\right]^2$$

$$= \sum\sum\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2 + \hat{\beta}\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)^2 - 2\hat{\beta}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..}\right)\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)$$

$$= E_{yy} + \frac{E_{xy}^2}{E_{xx}^2} * E_{xx} - 2\frac{E_{xy}}{E_{xx}} * E_{xy}$$

$$SSE = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

Degrees of freedom for SSE = d.f. for total - d.f. for treatment-d.f. for blocks- d.f. for β

Degrees of freedom for SSE = (n-1) – (p-1) – (q-1) -1 = (p-1)(q-1) -1

Under null hypothesis H_0: All treatments effects are equal i.e., $\alpha_i = 0$

∴ The model 1 becomes:

$$y_{ij} = \mu + \theta_j + \beta'\left(x_{ij} - \bar{x}_{..}\right) + e_{ij}^*$$

Thus, the restricted error sum of square under H_0 is given by:

$$SSE^* = \sum_i \sum_j\left[y_{ij} - \mu - \theta_j - \beta'\left(x_{ij} - \bar{x}_{..}\right)\right]^2$$

For finding the estimates of μ, θ_j *and* β' we need.

$$\frac{\partial SSE^*}{\partial \mu} = 0, \quad \frac{\partial SSE^*}{\partial \theta_j} = 0, \quad \frac{\partial SSE^*}{\partial \beta'} = 0$$

Now $\frac{\partial SSE^*}{\partial \mu} = 0$ from this we get:

$$\hat{\mu} = \bar{y}_{..}$$

Again

$$\frac{\partial SSE^*}{\partial \theta_j} = 2\sum_j[y_{ij} - \hat{\mu} - \hat{\theta}_j - \hat{\beta}'(x_{ij} - \bar{x}_{..})] = 0$$

$$\Rightarrow \sum_j y_{ij} - \sum_i \bar{y}_{..} - \sum_i \hat{\theta}_j - \hat{\beta}' \sum_i\left(x_{ij} - \bar{x}_{..}\right) = 0$$

$$\Rightarrow \hat{\theta}_j = \left(\bar{y}_{.j} - \bar{y}_{..}\right) - \beta'(\bar{x}_{.j} - \bar{x}_{..})$$

$\Rightarrow \hat{\theta}_j = (\bar{y}_{.j} - \bar{y}_{..}) - \beta'(\bar{x}_{.j} - \bar{x}_{..})$

Again,

$\frac{\partial SSE^*}{\partial \beta'} = -2 \sum_i \sum_j [y_{ij} - \hat{\mu} - \hat{\theta}_j - \hat{\beta}'(x_{ij} - \bar{x}_{..}](x_{ij} - \bar{x}_{..}) = 0$

$\Rightarrow \sum_i \sum_j [y_{ij} - \bar{y}_{..} - (\bar{y}_{.j} - \bar{y}_{..}) + \beta'(x_{.j} - \bar{x}_{..}) - \hat{\beta}'(x_{ij} - \bar{x}_{.j})](x_{ij} - \bar{x}_{..}) = 0$

$\Rightarrow \sum_i \sum_j [(y_{ij} - y_{.j}) - \hat{\beta}'(x_{ij} - \bar{x}_{.j})]\ (x_{ij} - \bar{x}_{..}) = 0$

$\Rightarrow \sum_i \sum_j [(y_{ij} - \bar{y}_{.j}) - \hat{\beta}'(x_{ij} - \bar{x}_{.j})](x_{ij} - \bar{x}_{.j} + \bar{x}_{.j} - \bar{x}_{..}) = 0$

$\Rightarrow \sum_i \sum_j (y_{ij} - \bar{y}_{.j})(x_{ij} - \bar{x}_{.j}) + \sum_i \sum_j (y_{ij} - \bar{y}_{.j})(\bar{x}_{.j} - \bar{x}_{..}) - \hat{\beta}' \sum_i \sum_j (x_{ij} - x_{.j})^2 -$

$\hat{\beta}' \sum_i \sum_j (x_{ij} - \bar{x}_{.j})(\bar{x}_{.j} - \bar{x}_{..}) = 0$

$\Rightarrow \sum_i \sum_j (y_{ij} - \bar{y}_{.j})(x_{ij} - \bar{x}_{.j}) - \hat{\beta}' \sum_i \sum_j (x_{ij} - \bar{x}_{.j})^2 = 0$

$\Rightarrow \hat{\beta}' = \frac{\sum\sum(y_{ij} - \bar{y}_{.j})(x_{ij} - \bar{x}_j)}{\sum_i \sum_j (x_{ij} - \bar{x}_{.j})^2} = \frac{E'_{xy}}{E'_{xx}}$

Now substituting the values of $\hat{\mu}$, $\hat{\theta}_J$ and $\hat{\beta}$ in the restricted SSE^*,we get

$SSE^* = \sum_i \sum_j [(y_{(ij)} - \bar{y}_{..} - (\overline{y_{.J}} - \bar{y}..) + \hat{\beta}(\overline{x_{.J}} - \bar{x}..) - \hat{\beta}(x_{(ij)} - \bar{x}..)]^2$

$= \sum_i \sum_j [(y_{(ij)} - \overline{y_{.J}}) - \hat{\beta}(x_{(ij)} - \bar{x}.j)]^2$

$= \sum_i \sum_j (y_{(ij)} - \overline{y_{.J}})^2 + \hat{\beta}^2 (x_{(ij)} - \bar{x}.j)^2 - 2\hat{\beta} \sum_i \sum_j [(y_{(ij)} - \overline{y_{.J}})(x_{(ij)} - \bar{x}.j)$

$= E'_{yy} + \frac{E'^2{}_{xy}}{E'^2{}_{xx}} * E'_{yy} - 2\frac{E'_{xy}}{E'_{xx}} * E'_{xy}$

$SSE^* = E'_{yy} - \frac{E'^2{}_{xy}}{E_{xx}}$

Degrees of freedom for $SSE^* = d.f$ for total -d.f. for blocks-d.f. for β

$$= (pq - 1) - (q - 1) - 1 = pq - q - 1$$

∴ The Adjusted Sum of Square for Treatment is $SST = SSE * - SSE$

d.f. for SST = d.f. for SSE^* − d.f. for SSE

$= pq - q - 1 - (p - 1)(q - 1) + 1$

$= pq - q - pq + p + q - 1$

$= p - 1$

MST $= \frac{SST}{p-1} = \frac{SSE*-SSE}{p-1}$

And $MSE = \frac{SSE}{(p-1)(q-1)-1}$

Now to test our null hypothesis H_o ,we can make a test statistic given as:

$$F = \frac{MST}{MSE} \sim F^{\alpha}\{(p-1), (p-1)(q-1)-1)$$

If $F > F^{\alpha}\{(p-1), (p-1)(q-1)-1)$,then our null hypothesis is rejected and we conclude that the treatments are effective(different), otherwise we accept the null and we conclude that there is no significant difference among the treatments.

ANOVA Table for Two-Way Classification (RBD Layout)

Sources of Variance	***Degree of freedom***	***Sum of Square***			***Estimate of β***	***Adjusted*** $SS_{(yy)}$	***Adjusted Degree of freedom***
		$SS_{(xx)}$	$SP_{(xy)}$	$SS_{(yy)}$			
Blocks	q-1	$B_{(xx)}$	$B_{(xy)}$	$B_{(yy)}$			
Treatment	p-1	$T_{(xx)}$	$T_{(xy)}$	$T_{(yy)}$		SSE	
Error	(p-1)(q-1)	$E_{(xx)}$	$E_{(xy)}$	$E_{(yy)}$	$\hat{\beta} = \frac{E_{(xy)}}{E_{(xx)}}$	SSE	(p-1)(q-1)-1
Treatment + Error	$q(p-1)$	$E'_{(xx)}$	$E'_{(xy)}$	$E'_{(yy)}$	$\hat{\beta} = \frac{E'_{(xy)}}{E'_{(xx)}}$	SSE*	$q(p-1) - 1$
Difference						SSE*- SSE	$p-1$

Example: In an experiment on cotton with 5 manurial treatments, it was observed that the number of plants per plot is varying from plot to plot. The yields of cotton along with number of plants per plot are given in the table below. Analyse the yield data removing the effect of variation in plant population on the yield by analysis of covariance technique and draw your conclusions. The design adopted was a RBD with four replications.

Treatments: 5 Levels of Nitrogen: $N_0 = 0$, $N_1 = 20$, $N_2 = 40$, $N_3 = 60$, $N_4 = 80$ kg/ha

Yield of Cotton (Number of plants) per plot

Replicate-I	N_1 12.0 (24)	N_0 10.5 (30)	N_4 27.0 (30)	N_2 16.5 (28)	N_3 25.0 (35)
Replicate-II	N_3 26.0 (40)	N_2 20.0 (25)	N_0 12.0 (25)	N_4 26.0 (22)	N_1 15.5 (28)
Replicate-III	N_2 22.0 (32)	N_4 30.0 (35)	N_3 20.0 (24)	N_1 20.0 (35)	N_0 14.5 (30)
Replicate 14	N_1 19.0 (26)	N_3 18.5 (16)	N_0 8.5 (24)	N_4 29.0 (30)	N_2 25.0 (35)

Solution: We set up the following hypothesis:

Null hypothesis:

H_{oT}: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5$, i.e., the treatments are homogenous

H_{oR}: $b_1 = b_2 = b_3 = b_4$, i.e., the blocks or replicates are homogenous

Alternative hypothesis:

H_{1T}: At least two τ_i' are different

H_{1R}: At least two b_j' are different

We shall use the ANCOVA technique to test these hypotheses:

y: Yield of cotton per plot

x (Concomitant variable): Number of plants per plot

Calculation of Various Sum of Squares

Treatments	***Yield of cotton (in kg.) along with the number of plants per plot***									
	Replication-I		*Replication-II*		*Replication-III*		*Replication-IV*		*Total*	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
N_0	30.0	10.5	25.0	12.0	30.0	14.5	24.0	8.5	109.0	45.5
N_1	24.0	12.0	28.0	15.5	35.0	20.0	26.0	19.0	113.0	66.5
N_2	28.0	16.5	25.0	20.0	32.0	22.0	35.0	25.0	120.0	83.5

N_3	35.0	25.0	40.0	26.0	24.0	20.0	16.0	18.5	115.0	89.5
N_4	30.0	27.0	22.0	26.0	35.0	30.0	30.0	29.0	117.0	112.0
Total	147.0	91.0	140.0	99.5	156.0	106.5	131.0	100.0	574.0	397.0

In usual notations, we have n = pq = 5*4 = 20; G = Grand total of all the observations.

For x:

G(x) = 574.0; n = 20

$$Correction\ Factor\ (CF(x)) = \frac{[G(x)]^2}{n} = \frac{[574.0]^2}{20} = 16{,}473.80$$

$$RSS_{xx} = \sum_i \sum_j x_{ij}^2 = 30^2 + 25^2 + \ldots\ldots + 35^2 + 30^2 = 17{,}086.00$$

$$Total\ SS\ (SS_{xx}) = RSS_{xx} - CF(x) = 17086.00 - 16473.80 = 612.20$$

$$\mathrm{R_{xx}} = \mathrm{SS(Replications)} = \frac{147^2+140^2+156^2+131^2}{5} - \mathrm{CF(x)} = 16541.20 - 16473.80 = 67.40$$

$$\mathrm{T_{xx}} = \mathrm{SS(Treatments)} = \frac{109^2+113^2+120^2+115^2+117^2}{4} - \mathrm{CF(x)} = 16491 - 16473.80 = 17.20$$

$$\mathrm{E_{xx}} = \mathrm{SS(Error)} = \mathrm{S_{xx}} - \mathrm{R_{xx}} - \mathrm{T_{xx}} = 612.20 - 67.40 - 17.20 = 527.60$$

For y:

G(y) = 397.0; n = 20

$$Correction\ Factor\ (CF(y)) = \frac{[G(y)]^2}{n} = \frac{[397.0]^2}{20} = 7{,}880.45$$

$$RSS_{yy} = \sum_i \sum_j y_{ij}^2 = 10.5^2 + 12.0^2 + \ldots\ldots + 30^2 + 29^2 = 8{,}652.50$$

$$Total\ SS\ (SS_{yy}) = RSS_{yy} - CF(y) = 8{,}652.50 - 7{,}880.45 = 772.05$$

$$\mathrm{R_{yy}} = \mathrm{SS(Replications)} = \frac{91^2+99.5^2+106.5^2+100^2}{5} - \mathrm{CF(y)} = 7904.70 - 7880.45 = 24.25$$

$$\mathrm{T_{yy}} = \mathrm{SS(Treatments)} = \frac{45.5^2+66.5^2+83.5^2+89.5^2+112^2}{4} - \mathrm{CF(y)} = 8504.75 - 7880.45$$

$$= 624.30$$

$$\mathrm{E_{yy}} = \mathrm{SS(Error)} = \mathrm{S_{yy}} - \mathrm{R_{yy}} - \mathrm{T_{yy}} = 772.05 - 24.25 - 624.30 = 123.50$$

For product xy:

G(x) = 574.0, G(y) = 397.0; n = 20

$Correction\ Factor\ (CF(xy)) = \frac{G(x)G(y)}{n} = \frac{574.0\ x\ 397.0}{20} = 11{,}393.90$

$RSS\ (Products) = RSS_{xy} = \sum\sum xy = (30\ x\ 10.5) + (25\ x\ 12) + \ldots\ldots + (30\ x\ 29)$

$= 11{,}704.00$

$Total\ SS\ (SP_{xy}) = RSS_{xy}\ -\ CF(xy)\ =\ 11{,}704.00\ -\ 11{,}393.90\ =\ 310.10$

$\mathrm{R_{xy}} = \mathrm{SP}_{xy}(\mathrm{Replications}) = \frac{(147*91)+(140*99.5)+(156*106.5)+(131*100)}{5} - \mathrm{CF(xy)}$

$= 11{,}404.20\ -\ 11{,}393.90 = 10.30$

$\mathrm{T_{xy}} = \mathrm{SP}_{xy}(\mathrm{Treatments}) = \frac{(109*45.5)+(113*66.5)+(120*83.5)+(115*89.5)+(117*112)}{4} - \mathrm{CF(xy)}$

$= 11{,}472.63\ -\ 11{,}393.90 = 78.73$

$\mathrm{E_{xy}} =\ \mathrm{SP}_{xy}(\mathrm{Error})\ =\ Total\ SS\ (SP_{xy})\ -\ \mathrm{R_{xy}} - \ \mathrm{T_{xy}} = 310.10\ -\ 10.30\ -\ 78.73\ =$

221.07

Sum of Squares and Products

Sources of Variance	***Degree of freedom***	***Sum of Square***			$\boldsymbol{MS}_{(yy)}$	$\boldsymbol{F}_{(yy)}$
		$\boldsymbol{SS}_{(xx)}$	$\boldsymbol{SP}_{(xy)}$	$\boldsymbol{SS}_{(yy)}$		
Replications	4 – 1 = 3	$R_{(xx)} =$ 67.40	$R_{(xy)} =$ 10.30	$R_{(yy)} =$ 24.25	$\frac{24.25}{3} =$ 8.08	
Treatments	5 – 1 = 4	$T_{(xx)} =$ 17.20	$T_{(xy)} =$ 78.73	$T_{(yy)} =$ 624.30	$\frac{624.30}{4} =$ 156.08	$\frac{156.08}{10.29} =$ 15.17
Error	3*4 = 12	$E_{(xx)} =$ 527.60	$E_{(xy)} =$ 221.07	$E_{(yy)} =$ 123.50	$\frac{123.50}{12} =$ 10.29	
Total	20 – 1 = 19	$SS_{(xx)} =$ 612.20	$SS_{(xy)}$ = 310.10	$SS_{(yy)}$ = 772.05		

We now adjust for variation in yield (y) from plot to plot for the linear (regression) effect of the number of plants (x) per plot. An estimate of the coefficient of regression (β) of y on x is given by:

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}} = \frac{221.07}{527.60} = 0.42$$

The adjusted (corrected) error sum of squares for y, adjusted for this linear effect is given by:

$$Adjusted\ Error\ SS\ for\ y\ =\ Adjusted\ (E_{yy}) = E_{yy}\ -\ \hat{\beta}E_{xy} =\ E_{yy}\ -\ \frac{{E_{xy}}^2}{E_{xx}}$$

$$= 123.50\ -\ \frac{221.07^2}{527.60} = 30.80$$

(OR: $\hat{\beta}E_{xy} =\ 0.4190\ *\ 221.07\ =\ 92.63$)

The estimation of $\hat{\beta}$ results in loss of 1 d.f. for error sum of squares, which now becomes 12-1=11.

Next, the variation in treatments is also to be adjusted for variation in x. For this, we prepare the following table for (Treatments + Error) sum of squares.

Sum of Squares and Sum of Products for (Treatments + Error)

Source of Variation	***SS(x^2)***	***SP(xy)***	***SS(y^2)***
Treatments	T_{xx} = 17.20	T_{xy} = 78.73	T_{yy} = 624.30
Error	E_{xx} = 527.60	E_{xy} = 221.07	E_{yy} = 123.50
Treatments + Error	E'_{xx} = 544.80	E'_{xy} = 299.80	E'_{yy} = 747.80

$E'_{xx} = T_{xx} + E_{xx}, E'_{xy} = T_{xy} + E_{xy}, E'_{yy} = T_{yy} + E_{yy}$

The Sum of Square for (Treatments + Error) for y is adjusted for linear (regression) effect of x on y exactly similarly as the error sum of square and is given by:

SSE* = Adjusted S.S. for (Treatments + Error) for y = $E'_{yy}\ -\ \frac{(E'_{xy})^2}{E'_{xx}}$

$$= 747.80\ - \frac{(299.80)^2}{544.80}\ = 582.82$$

Finally, the Treatments S.S. for y adjusted for the linear effect of x on y is given by:

Adjusted (Treatment S.S.) for y = SSE* - Adjusted (E_{yy})

= Adjusted (Treatment + Error) S.S. for y – Adjusted Error S.S. for y

= 582.82 – 30.87 = 551.95

Adjusted Analysis of Variance Table

Source of Variation	*Degree of Freedom*	*Sum of Squares*	*Mean Sum of Squares*	*Variance Ratio*
Treatments	4	551.95	137.99	F_T = 49.11
Error	12 – 1 = 11	30.87	2.81	
Treatment + Error	16 – 1 = 15	582.82		

$Tabulated\ F_{(4,11)}(0.05) = 3.36$

Conclusion: Since the calculated value of F_T = 49.11 is much greater than the tabulated (critical) value. It is highly significant. Hence, we reject the null hypothesis of equality of treatment means and conclude that the treatments differ significantly as regards their effect on increase of yield of cotton. Moreover, from the sum of squares table, we conclude that the treatment N_4 is the most effective, followed by N_3, N_2, N_1 and N_0 respectively.

3.6 Self-Assessment Exercise

1. An experiment was conducted to judge the effectiveness three drugs in reducing blood pressure for three different groups of people. The data of the amount of blood pressure reduction (in millimeters of mercury) is given as follows:

Group	*Drug*		
	X	*Y*	*Z*
A	14 15	10 9	11 11
B	12 11	7 8	10 11
C	10 11	11 11	8 7

Answer the following questions taking significant level of 5%

a) Do the drugs act differently?

b) Are the different groups of people affected differently?

c) Is the interaction term significant?

2. A manufacturer wishes to determine the effectiveness of four types of machines (A, B, C and D) in the production of bolts. To accumulate this, the numbers of defective bolts produced for each of two shifts in the results are shown in the following table:

Machine	*First Shift*					*Second Shift*				
	M	*T*	*W*	*Th*	*F*	*M*	*T*	*W*	*Th*	*F*
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

Perform an analysis of variance to determine at 5% level of significance, whether there is a difference between the machines and between the shifts.

3. Explain the process of assessing whether there is a significant interaction effect between the two factors in a two-way layout with one observation per cell.
4. What do you understand by "Analysis of Covariance"? Illustrate with suitable examples.
5. Derive the Analysis of Covariance for a one-way layout (with one concomitant variable only).

3.7 Summary

This unit provides an overview of the analysis of variance in two-way classified data with m-observations per cell, how to perform Tukey's Test for Non-Additivity for Two-way layout with one observation per cell and conduct Analysis of Covariance (ANCOVA) for one-way and two-way classified data.

3.8 References

- Gupta, S.C. and Kapoor, V.K. (2008). *Fundamentals of Applied Statistics (4th ed.)*, Sultan Chand and Sons.
- Mukhopadhyay, P (2011). *Applied Statistics* (2nd edition revised reprint), Books & Allied (P) Ltd.

3.9 Further Reading

- Bapat, R. (2012). *Linear Algebra and Linear Models*. Springer.
- Gun, A.M., Gupta, M.K. and Dasgupta, B. (2013). *Fundamentals of Statistics Vol.1 & II*. World Press.

- Hogg, R. V., McKean, J. W., and Craig, A. T. (2020). *Introduction to Mathematical Statistics*. Pearson.
- Mood, A. H., Bose, D. C., & Graybill, F. A. (1984). *Introduction to the theory of Statistics*. McGraw-Hill.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N Linear Models and Design of Experiment

Block: 2 Design of Experiment

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Shambhavi Mishra **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Editor**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N **LINEAR MODEL & DESIGN OF EXPERIMENT**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Blocks & Units Introduction

The ***Block - 2 –Design of Experiment*** is the second block of said SLM with three units.

In ***Unit – 4 – Basic Designs,*** is being introduced the Terminology and basic Principles of Design, CRD, RBD and LSD, analysis with missing observations.

In ***Unit – 5 – Factorial Experiments*** is discussed with 2^3, 2^n, 3^2 and 3^3 factorial experiments with its analysis.

In ***Unit – 6 – Confounding*** has been introduced, Orthogonality, Complete and Partial confounding, construction of confounded factorial experiments.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT: 4 BASIC DESIGNS

Structure

4.1 Introduction

In 1935 R.A. Fisher laid the foundation of the subject called Design of Experiments. Applications of this theory are found in laboratories and research in natural sciences, engineering and in almost all branches of social sciences. The subject matter of the design of experiment consists of:

1) Planning of the experiment,
2) Obtaining relevant information from it regarding the statistical hypothesis under study, and
3) Making a statistical analysis of the data.

Experiment: It is a device or means of getting an answer to the question that the experimenter has in mind. Experiment can be classified into two categories: (a) Absolute and (b) Comparative.

(a) Absolute Experiments: This consists of determining the absolute value of some characteristics like finding the mean of a set of data or finding the correlation coefficients between two variables or finding the variability of a data *etc.*

(b) Comparative experiments: These are designed to compare the effect of two or more objects on some population characteristic, *e.g.,* comparison of different manures or fertilizers, different diets or medicines in a medical experiment or standards of teaching in different educational institutions *etc.*

Treatments: Various objects of comparison in a comparative experiment are called as different treatments. For example, in field experiment different fertilizers or different varieties of crop *etc;* in an experiment regarding comparison of standards of teaching in different institutions, the different institutions will be the treatments, while in an experiment concerning effect of different drugs on patients suffering from certain disease, the treatment will be the different drugs used to cure them.

Experimental Unit: The smallest division of the experimental material to which the treatment is applied and on which the observation of the variable under study is made is known as an experimental unit. For example, in field experiment, the plot or land is the experimental unit. In other experiments it may be a patient in a hospital, students of a particular class of an institution, a lump of dough or a batch of seeds *etc.*

Blocks: In agricultural experiments most of the times we divide the whole experimental unit (field) into relatively homogeneous subgroups or strata. These strata which are more uniform amongst themselves than the field as a whole are known as blocks.

Yield: The measurement of the variable under study on different experimental units are termed as yields.

Experimental error: The chance or non-assignable cause of variation is termed as experimental error.

Replication: It is the execution of an experiment more than once, *i.e.* repetition of treatments under investigation.

Precision: The reciprocal of the variance of the mean is termed as the precision or the amount of information of a design. Thus, for an experiment replicated *r* times if $\bar{x}$ denotes the mean of the observed values of yield, then, $V(\bar{x}) = \frac{\sigma^2}{r}$, where σ^2 is the variance of each individual observations or error variance per unit. Then, $Precision = \frac{1}{V(\bar{x})} = \frac{r}{\sigma^2}$.

Efficiency of a Design: Consider two designs D_1 and D_2 with error variances σ_1^2 and σ_2^2 and replications r_1 and r_2 respectively. Then, the precision of $D_1 = \frac{r_1}{\sigma_1^2}$ and the precision of $D_2 = \frac{r_2}{\sigma_2^2}$. Efficiency of the design D_1 with respect to the design D_2 is defined as:

$E = \frac{Precision\ of\ D_1}{Precision\ of\ D_2} = \frac{r_1/\sigma_1^2}{r_2/\sigma_2^2}$.

If E = 1, then both the designs D_1 and D_2 are equally efficient.

If $E > 1$ (or $E < 1$) then D_1 is said to be more (or less) efficient than D_2.

Uniformity trials: By uniformity trials we mean a trial in which the experimental material is divided into small units and the same treatment is applied on each of the units and their yields are recorded. By doing so we can have an idea about the uniformity of the experimental material, for example in case of field experiment, the variation of the fertility gradient of the land can be identified which will be helpful in the formation of the blocks.

4.2 Objectives

After going through this unit, you should be able to:

- Understand the principles of design of experiments
- Apply the Completely Randomized Design (CRD), Randomized Block Design (RBD), and Latin Square Design (LSD)
- Estimate the missing value(s) in RBD and LSD.

4.3 Principles of Design of Experiments

According to Prof. R.A. Fisher, there are three basic principles of design of experiments. They are:

Replication: As said earlier, replication is the repetition of the treatments under investigation. This is done to average out the influence of chance factors on different experimental units. Thus, repetition of treatments results in more reliable estimate than a single observation.

Randomisation: It is a process of assigning treatments to various experimental units in a pure chance manner. This gives each treatment equal chance of showing its worth. The purpose of randomness is to assume that the sources of variation, not controlled by the experiment, *i.e.*, chance variation operate randomly so that the average effect of it on any group or unit is zero.

Local Control: The process of dividing a heterogeneous experimental material into homogeneous groups or blocks is known as local control. For example, in an agricultural field experiment, the whole experimental area (field) is divided into groups (blocks) row-wise or column-wise or both according to the fertility gradient of the soil such that the variation within each block is minimum and between blocks is maximum. The treatments are allocated within each block at random.

4.4 Completely Randomized Design (CRD)

This design is based on the principle of randomisation and replication. In this design treatments are allocated at random to the experimental units over the entire experimental material.

4.4.1 Layout

Suppose there are k treatments and each treatment is replicated $r_1, r_2, \cdots, r_k$ number of times, then the whole experimental material is divided into $n = \sum_{i=1}^{k} r_i$ units and all the $\sum_{i=1}^{k} r_i$

treatments are allocated to these *n* units completely at random. For example, suppose there are 3 treatments t_1, t_2 and t_3 which are replicated 2,3 and 5 times. Then the experimental material is divided into 2+3+5= 10 units and the treatments are allocated at random to these 10 units as follows:

t_3	t_2	t_1	t_2	t_2
t_3	t_3	t_1	t_3	t_3

The analysis of variance of a CRD is exactly as that of a one-way layout.

Assumptions

The statistical analysis of this layout is based on the following assumptions:

1) All the observations are mutually independent.
2) Different effects are additive in nature.
3) e_{ij}'s are i.i.d. random variables each following $N(0, \sigma_e^2)$; $j = 1,2,\cdots,n_i, i = 1,2,\cdots,k$.

4.4.2 Analysis

Suppose there are *k* treatments and each treatment is replicated $r_1, r_2, \cdots, r_k$ number of times, then the whole experimental material is divided into $n = \sum_{i=1}^{k} r_i$ units and all the $\sum_{i=1}^{k} r_i$ treatments are allocated to these *n* units completely at random. Let these *n* observations be denoted as y_{ij},$(i = 1,2,\cdots,k; j = 1,2,\cdots,r_i)$. Then the observation table is as follows:

Treatments	***Observations***	***Total***	***Mean***
1	y_{11} y_{12} $\cdots\cdots\cdots\cdots\cdots\cdots$ y_{1r_1}	$T_{1.} = \sum_{j=1}^{r_1} y_{1j}$	$\bar{y}_{1.} = \frac{T_{1.}}{r_1}$
2	y_{21} y_{22} $\cdots\cdots\cdots\cdots\cdots\cdots$ y_{2r_2}	$T_{2.} = \sum_{j=1}^{r_2} y_{2j}$	$\bar{y}_{2.} = \frac{T_{2.}}{r_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1} y_{i2} $\cdots\cdots\cdots\cdots\cdots\cdots$ y_{ir_i}	$T_{i.} = \sum_{j=1}^{r_i} y_{ij}$	$\bar{y}_{i.} = \frac{T_{i.}}{r_i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

k	y_{k1} y_{k2} $\cdots\cdots\cdots\cdots\cdots\cdots$ y_{kr_k}	$T_{k.} = \sum_{j=1}^{r_k} y_{kj}$	$\bar{y}_{k.} = \frac{T_{k.}}{r_k}$
Total		$T_{..} = \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}$	$\bar{y}_{..} = \frac{T_{..}}{n}$

The mathematical model is given by

$y_{ij} = \mu_i + e_{ij}; j = 1,2,\cdots,r_i, i = 1,2,\cdots,k,$

where μ_i is the average effect of the i^{th}treatment which can be split as:

$\mu_i = \mu + \mu_i - \mu = \mu + \alpha_i$ with $\alpha_i = \mu_i - \mu$, $i = 1,2,\cdots,k$ and $\mu = \frac{1}{n}\sum_{i=1}^{k} r_i\mu_i$.

Hence,

$y_{ij} = \mu + \alpha_i + e_{ij}; j = 1,2,\cdots,r_i, i = 1,2,\cdots,k;$ (1)

Where,

y_{ij} is the j^{th} observation of i^{th} treatment; $j = 1,2,\cdots,r_i, i = 1,2,\cdots,k,$

μ is the general mean effect,

α_i is the additive effect due to i^{th}treatment

e_{ij} is the error effect due to chance and these are assumed to be *i.i.d.* random variables each following $N(0,\sigma_e^2); j = 1,2,\cdots,r_i, i = 1,2,\cdots,k.$

The side condition is $\sum_{i=1}^{k} r_i\alpha_i = \sum_{i=1}^{k} r_i(\mu_i - \mu) = n\mu - n\mu = 0$.

The null hypothesis to be tested is:

H₀: The groups do not differ significantly or have no additive effect due to different groups. In other words, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$.

Summing (1) over *j* and dividing by r_i, we get:

$\bar{y}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k,$ (2)

where $\bar{e}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} e_{ij}$ are *i.i.d.* random variables each distributed as $N(0,\sigma_e^2/n_i)$.

Summing (1) over *i* and j and dividing by n, we get

$\bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij} = \mu + \bar{e}_{..} = \mu + \bar{e}_{..}$ (3)

where $\bar{e}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$ are *i.i.d.* random variables each distributed as $N(0,\sigma_e^2/n)$.

Now the total variation in each observation is given by the total sum of squares as:

$T.S.S. = \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(\bar{y}_{i.} - \bar{y}_{..} + y_{ij} - \bar{y}_{i.}\right)^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{r_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$

$= \sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2.$

Or T.S.S. = *S.S.T.* + *S.S.E*,

Where, *T.S.S* = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{..}\right)^2$;*S.S.T* = Sum of squares due to treatments = $\sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..})^2$; and *S.S.E* = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$.

Degrees of Freedom

T.S.S = Total sum of squares = $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{..}\right)^2$ is computed from *n* quantities of the form $\left(y_{ij} - \bar{y}_{..}\right)$ with one constraint $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{..}\right) = 0$. Hence, *T.S.S* will have *n* – 1 degrees of freedom.

S.S.T = Sum of squares due to treatments = $\sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..})^2$ is computed from *k* quantities of the form $(\bar{y}_{i.} - \bar{y}_{..})$ with one constraint $\sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Hence, *S.S.G* will have *k* – 1 degrees of freedom.

S.S.E = Sum of squares due to error or residuals = $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$ is computed from *n* quantities of the form $\left(y_{ij} - \bar{y}_{i.}\right)$ with *k* constraints $\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right) = 0$. Hence, *S.S.E* will have *n* – *k* degrees of freedom.

Mean Sum of Squares

The sum of squares divided by its degrees of freedom gives the corresponding mean sum of squares. Thus,

Mean Sum of Squares due to Treatments (M.S.T.) = $\frac{S.S.T.}{k-1}$.

Mean Sum of Squares due to Error (M.S.E.) = $\frac{S.S.E.}{n-k}$

4.4.3 Least Square Estimates

In the mathematical model (1), μ and $\alpha_i, i = 1,2,\cdots,k$ are the unknown parameters which have to be estimated by the principle of least squares. Hence, we consider the sum of squares due to errors, which is given by:

$$S.S.E = \sum_{i=1}^{k}\sum_{j=1}^{r_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \mu - \alpha_i\right)^2. \quad (4)$$

Differentiating (4) with respect to μ and α_i and equating to zero individually, we get

$constant\ w.r.t.x \Rightarrow -2\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij} = n\mu + \sum_{i=1}^{k} r_i\alpha_i = n\mu \qquad [\because \sum_{i=1}^{k} r_i\alpha_i = 0 \text{ by side condition.}]$$

Hence, the estimate of μ is given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij} = \bar{y}_{..}$$

$$\frac{dS.S.E}{d\alpha_i} = 0 \Rightarrow -2\sum_{j=1}^{r_i}\left(y_{ij} - \mu - \alpha_i\right) = 0,\ i = 1,2,\cdots,k$$

$$\Rightarrow \sum_{j=1}^{r_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

$$\Rightarrow \sum_{j=1}^{r_i} y_{ij} = r_i\mu + r_i\alpha_i$$

$$\Rightarrow \hat{\alpha}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} y_{ij} - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$$

Variance of the estimates

We have $\hat{\mu} = \bar{y}_{..}$ and $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$

$$V(\hat{\mu}) = E[\bar{y}_{..} - E(\bar{y}_{..})]^2 = E[\mu + \bar{e}_{..} - \mu]^2 = E[\bar{e}_{..}]^2 = E(\bar{e}_{..}^2)$$

$$= V(\bar{e}_{..}) = \frac{\sigma_e^2}{n}$$

Also, we have $\hat{\alpha}_i - E(\hat{\alpha}_i) = \bar{y}_{i.} - \bar{y}_{..} - E(\bar{y}_{i.} - \bar{y}_{..})$

$$= \mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..} - E(\mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..})$$

$$= \alpha_i + \bar{e}_{i.} - \bar{e}_{..} - \alpha_i$$

$$= \bar{e}_{i.} - \bar{e}_{..}$$

Hence, $V(\hat{\alpha}_i) = E[\bar{e}_{i.} - \bar{e}_{..}]^2 = E[\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}]$

$= E(\bar{e}_{i.}^2) + E(\bar{e}_{..}^2) - 2E(\bar{e}_{i.}\bar{e}_{..})$

Now, $E(\bar{e}_{i.}\bar{e}_{..}) = E(\frac{1}{r_i}\sum_{j=1}^{r_i} e_{ij} \frac{1}{kr_i}\sum_{i=1}^{k}\sum_{j=1}^{r_i} e_{ij})$

$= \frac{1}{kr_i^2} E[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ir_i}^2] + \frac{1}{kr_i^2} E\left[\sum_{j=1}^{n_i} e_{ij} \sum_{h\neq i=1}^{k}(e_{h1} + \cdots + e_{hr_i})\right]$

$= \frac{1}{kr_i^2} E[e_{i1}^2 + e_{i2}^2 + \cdots + e_{ir_i}^2]$ since $E(e_{ij}e_{hj}) = 0$ for $h \neq i$;

$= \frac{1}{kr_i^2}\sum_{j=1}^{r_i} E(e_{ij}^2) = \frac{1}{kr_i^2}\sum_{j=1}^{r_i} V(e_{ij}) = \frac{1}{kr_i^2} r_i\sigma_e^2 = \frac{\sigma_e^2}{kr_i}$.

Hence, $V(\hat{\alpha}_i) = \frac{\sigma_e^2}{r_i} + \frac{\sigma_e^2}{n} - 2\frac{\sigma_e^2}{kn_i} = \frac{\sigma_e^2}{r_i}\left(1 - \frac{2}{k}\right) + \frac{\sigma_e^2}{n}$

In particular if all treatments are repeated the same number of times, say equal to r, *i.e.*, if

$r_i = r, \forall\, i = 1,2,\cdots,k$, then $n = rk$ and

$V(\hat{\alpha}_i) = \frac{\sigma_e^2}{r}\left(1 - \frac{2}{k}\right) + \frac{\sigma_e^2}{rk} = \frac{\sigma_e^2}{r}\left(1 - \frac{2}{k} + \frac{1}{k}\right) = \frac{(k-1)\sigma_e^2}{rk}$

Expectation of Sum of Squares

We have $y_{ij} = \mu + \alpha_i + e_{ij}; j = 1,2,\cdots,r_i, i = 1,2,\cdots,k;$

$\bar{y}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} y_{ij} = \mu + \alpha_i + \bar{e}_{i.}, \forall\, i = 1,2,\cdots,k$, and

$\bar{y}_{..} = \mu + \bar{e}_{..}$,

Then:

$E(y_{ij}^2) = E(\mu^2 + \alpha_i^2 + e_{ij}^2 + 2\mu\alpha_i + 2\mu e_{ij} + 2\alpha_i e_{ij})$

$= E(\mu^2) + E(\alpha_i^2) + E(e_{ij}^2) + 2\mu E(\alpha_i) + 2\mu E(e_{ij}) + 2E(\alpha_i)E(e_{ij})$

$= \mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i$

$E(\bar{y}_{i.}^2) = E(\mu^2 + \alpha_i^2 + \bar{e}_{i.}^2 + 2\mu\alpha_i + 2\mu\bar{e}_{i.} + 2\alpha_i\bar{e}_{i.})$

$= E(\mu^2) + E(\alpha_i^2) + E(\bar{e}_{i.}^2) + 2\mu E(\alpha_i) + 2\mu E(\bar{e}_{i.}) + 2E(\alpha_i)E(\bar{e}_{i.})$

$= \mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{r_i} + 2\mu\alpha_i$

$E(\bar{y}_{..}^2) = E(\mu^2 + \bar{e}_{..}^2 + 2\mu\bar{e}_{..})$

$$= E(\mu^2) + E(\bar{e}_{..}^2) + 2\mu E(\bar{e}_{..}) = \mu^2 + \frac{\sigma_e^2}{n}$$

$$E(S.S.G.) = E\{\sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..})^2\}$$

$$= E\{\sum_{i=1}^{k} r_i\bar{y}_{i.}^2 - n\bar{y}_{..}^2\}$$

$$= \sum_{i=1}^{k} r_i E(\bar{y}_{i.}^2) - nE(\bar{y}_{..}^2)$$

$$= \sum_{i=1}^{k} r_i \left(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{r_i} + 2\mu\alpha_i\right) - n(\mu^2 + \frac{\sigma_e^2}{n})$$

$$= n\mu^2 + \sum_{i=1}^{k} r_i\,\alpha_i^2 + k\,\sigma_e^2 + 2\mu\sum_{i=1}^{k} r_i\alpha_i - n\mu^2 - \sigma_e^2$$

$$= \sum_{i=1}^{k} r_i\,\alpha_i^2 + (k-1)\sigma_e^2$$

Or $E(M.S.G.) = E\left(\frac{S.S.G}{k-1}\right) = \frac{1}{(k-1)}\sum_{i=1}^{k} r_i\,\alpha_i^2 + \sigma_e^2$

Now $E(S.S.E.) = E\{\sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\}$

$$= E\{\sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}^2 - \sum_{i=1}^{k} r_i\bar{y}_{i.}^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{r_i} E(y_{ij}^2) - \sum_{i=1}^{k} r_i E(\bar{y}_{i.}^2)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{r_i}(\mu^2 + \alpha_i^2 + \sigma_e^2 + 2\mu\alpha_i) - \sum_{i=1}^{k} r_i(\mu^2 + \alpha_i^2 + \frac{\sigma_e^2}{n_i} + 2\mu\alpha_i)$$

$$= n\mu^2 + \sum_{i=1}^{k} n_i\,\alpha_i^2 + n\sigma_e^2 + 2\mu\sum_{i=1}^{k} r_i\alpha_i - n\mu^2 - \sum_{i=1}^{k} r_i\,\alpha_i^2 - k\sigma_e^2 - 2\mu\sum_{i=1}^{k} r_i\alpha_i$$

$$= (n-k)\sigma_e^2.$$

Or $E(M.S.E.) = E\left(\frac{S.S.E}{n-k}\right) = \sigma_e^2$.

Thus, under H_0, $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. Hence, $E(M.S.T.) = \sigma_e^2 = E(M.S.E.)$.

Also, under H_0, *S.S.T.* follows a χ^2 distribution with $k - 1$ degrees of freedom and *S.S.E.* follows a χ^2 distribution with $n - k$ degrees of freedom.

Hence, for testing H_0, the test statistic is given by $F = \frac{S.S.T/(k-1)}{S.S.E./(n-k)} = \frac{M.S.T}{M.S.E}$ which will follow a central F distribution with $k - 1$ and $n - k$ degrees of freedom.

4.4.4 ANOVA Table

Sources of Variation	*Degrees of freedom*	*Sum of Squares*	*Mean Sum of Squares*	*Variance Ratio*
Treatments	$k-1$	$S.S.T. = \sum_{i=1}^{k} r_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$M.S.T = \frac{S.S.T.}{k-1}$	$F = \frac{M.S.T.}{M.S.E}$
Error	$n-k$	$S.S.E. = \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$	$M.S.E. = \frac{S.S.E}{n-k}$	
Total	$n-1$	$T.S.S. = \sum_{i=1}^{k}\sum_{j=1}^{r_i}\left(y_{ij} - \bar{y}_{..}\right)^2$		

If $F > F_{(k-1,n-k)}(\alpha)$ then H_0 is rejected at α% level of significance and we conclude that treatments differ significantly, H_0 may be accepted. If the calculated value of F is greater than the tabulated value of F at $k - 1$ and $n - k$ degrees of freedom, then reject the null hypothesis H_0.

If H_0 is rejected in that case, we proceed further to find out which of the treatment means differ significantly. For this, we find out the "Critical Difference (CD)", i.e., the least difference between any two means to be significant.

S.E. of the difference between any two-treatment means is:

$$C.D. = \sqrt{\frac{2*MSE}{r}} \text{ x } [t_{0.025}(for\ error\ d.f.)] = \sqrt{\frac{2*MSE}{r}} \text{ x } t_{d.f.}(0.025)$$

Where, r is the number of times a treatment is replicated.

If the critical difference is greater than the absolute mean difference between any two treatment means, then that pair of treatments differ significantly.

4.4.5 Advantages and Disadvantages of CRD

Advantages

CRD has several advantages which are given below:

1. It is easy to layout the design.
2. This design is very useful to conduct small experiment.
3. In certain type of laboratory experiments where the experimental units are homogeneous.
4. In this design we have complete flexibility on diving number of treatment and their replications. This procedure simplifies the analysis of data when observations of such experimental units or an entire treatment are missing.
5. The CRD provides the maximum degrees of freedom for the estimation of an experimental unit.

Disadvantage

1. In this design, the third principle of design of experiment i.e., principle of the local control is not used.
2. The design is rarely used in the field experiment because practically the plots are homogeneous.

Example: A set of data involving four "tropical feed stuffs A, B, C, D" tried on 20 chicks is given below. All the twenty chicks are treated alike in all respects except the feeding treatments and each feeding treatment is given to 5 chicks. Analyse the data.

Feed	***Gain in Weight***				
A	55	49	42	21	52
B	61	112	30	89	63
C	42	97	81	95	92
D	169	137	169	85	154

Solution:

Feed	***Gain in Weight***					***Total Ti***
A	55	49	42	21	52	219
B	61	112	30	89	63	355
C	42	97	81	95	92	407
D	169	137	169	85	154	714
	Grand Total					G = 1,695

The null hypothesis, H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k$ i.e., the treatment effects are same. In other words, all the treatments (A, B, C, D) are alike as regards their effect on increase in weight.

Alternative that all H_1: At least two of α'_s are different.

Raw S.S. (R.S.S.) = $\sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}2 = 55^2$+$49^2$ + ………. + 85^2 + 154^2 = 1,81,445

Correction factor (C.F.) = $\frac{G^2}{N}$ $\frac{(1695)^2}{N}$ = 1,43,651.25

Total S.S. (T.S.S.) = R.S.S. – C.F. = 1,81,445 – 1,43,651.25 = 37,793.75

Treatment S.S. (S.S.T.) = $\frac{{T_1}^2 + {T_2}^2 + {T_3}^2 + {T_4}^2}{5}$ – C.F.

= $\frac{47{,}961 + 1{,}26{,}025 + 1{,}65{,}649}{5}$ – 1,43,641.25 = 26,234.95

Error S.S. = TSS - SST= 37,793.75 - 26,234.95 = 11,558.80

ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$\boldsymbol{F_{Cal.}}$	$\boldsymbol{F_{Tab.}}$
Treatments	3	26,234.95	8744.98	$F_T = \frac{8744.98}{722.42} =$ 12.105*	$F_{0.05(3,16)} =$ 3.06
Error	16	11,558.80	722.42		
Total	19	37,793.75			

Here, $F_T > F_{0.05(3,16)}$, hence F_T is highly significant and we reject H_0 at 5% level of significance and conclude that the treatments A, B, C, and D differ significantly.

Since H_0 is rejected in this case, we proceed further to find out which of the treatment means differ significantly. For this, we find out the "Critical Difference (CD)", i.e., the least difference between any two means to be significant.

S.E. of the difference between any two-treatment means is =

$$C.D. = \sqrt{(2s_E^2/r)} \text{ x } [t_{0.025}(for\ error\ d.f.)] = \sqrt{\frac{2\ x\ 722.42}{5}}\ x\ 2.12 = 16.99\ x\ 2.12 = 36.018$$

The treatment mean effects, arranged in descending order of magnitude, are given as:

Treatment	***Mean gain in weight***	***Difference***	
D	142.8	142.8-81.4 = 61.4	37.6*
C	81.4	81.4-71.0 = 10.4	
B	71.0	71.0 – 43.8 = 24.2	
A	43.8		

Comparing these differences with the C.D., we find that:

i. Treatment D differs significantly from each of the treatments A, B and C,
ii. The treatments A and C also differ significantly, and
iii. All the remaining differences are not significant.

Conclusions - Treatments A, B, C and D are not alike. The highest treatment mean effect is 142.8 due to the feedstuff D. Hence if a choice is to be made among the four treatments A, B, C and D, treatment D is the best and most effective. Moreover, if a choice is to be made between A and C (which differ significantly), the treatment C is to be preferred since the average gain in weight due to treatment C is more than due to the treatment A. all other possible combination of treatment pairs is alike.

4.5 Randomized Block Design (RBD)

In this design all the three principles of design of experiment are used. If the experimental units are heterogeneous, the CRD is not used. In that case the variation among units affects the test of the significance of the treatment effects. The design which enables us to take care of the variability among unit is the RBD.

Suppose we wish to compare the effect of p treatment; each treatment being replicated an equal no. of finite say q. then we need n = pq experimental units. These units are not perhaps homogeneous. In RBD the first step is to divide units into q parts or more homogeneous group and each group or block we take as many units as here are treatments. Thus the no. of block is equal to the common replication number. The same techniques should be used to the units of the block. The variation in technique should be made between the blocks. Ex: In this field experiment if the fertility gradient is present, then it is advisable to place the blocks across the gradient in order to get homogeneous material for a block and to obtain the major difference among blocks. The second step is to assign the treatment at random to the unit of the block. This randomization has to be done for each block. However, in CRD randomization was restricted with in a homogeneous block. In this design each treatment will have same number of replications if we want to additional replication foe some treatment, then each of these may be applied to more than one unit in a block.

4.5.1 Layout

Suppose we have 5 treatments and each treatment are replicated 3 times, so we need 15 units, which are to be grouped to 3 blocks of 5 plots each. We numbered the treatments the unit in a block and by following any method of drawing a random sample we get a random permutation of digits from 1 to 5 say 4,3,1,5,2. For block 1^{st} then we apply treatment number 1 to unit 4, treatment number 2 to unit 3 and so on and finally treatment number 5 to unit 2, similarly we find another random permutation for block 2^{nd} and so on for other blocks.

4.5.2 Analysis

The analysis of this design is same as that of two-way classified data with one observation per cell. We use the following fixed effect model:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad (1)$$

Where y_{ij} is the value of j^{th} unit receiving the i^{th} treatment, μ is the general mean effect, α_i is the additional effects due to i^{th} treatment over general effects, β_j is the additional effect due to the j^th^ treatment over general effects and e_{ij} be the experimental error corresponding ij^th^ cell.

The model is based on the following assumptions:

i. All the observations are independently distributed.
ii. Different effects are additive in the nature.
iii. The error e_{ij} are independently and identically distributed with mean μ and variance σ_e^2 i.e., $eij \sim N(0, \sigma e2)$.

Here we are to test the equality or homogeneity of the different levels of factor A as well as different levels of factor B, thus our null hypothesis are:

H_{0A}= $\mu_{1.} = \mu_{2.} = \ldots\ldots\ldots = \mu_{p.} = \mu$ (=)$\alpha_1 = \alpha_2 = \ldots\ldots \alpha_p = 0$

H_{0B}= $\mu_{.1} = \mu_{.2} = \ldots\ldots\ldots = \mu_{.q} = \mu$ (=)$\beta_1 = \beta_2 = \ldots\ldots \beta_q = 0$

Against:

H_{1A}= Atleast two means are not equal.

H_{1B}= Atleast two means are not equal.

4.5.3 Least Square Estimates

To test the above hypothesis, we find the least square estimates by minimizing the residual sum of squares as:

$$S = \sum i \sum_j eij^2 = \sum i \sum_j (yij - \alpha i - \beta j)^2$$

The normal equations are:

$$\frac{\partial s}{\partial \mu} = 0, \frac{\partial s}{\partial \alpha i} = 0, \frac{\partial s}{\partial \beta j} = 0$$

And we get the following estimates:

$\hat{\mu} = \bar{y}..$

$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$

$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}..$

Now substituting these values of the estimates in linear model (1), we get:

$$y_{ij} = \bar{y}_{..} + (y_{i.} - \bar{y}_{..}) + (y_{.j} - \bar{y}_{..}) + (y_{ij} - y_{i.} - y_{.j} + \bar{y}_{..})$$

Now squaring both sides and summing over all observations, we get:

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j (y_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{.j} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.} - y_{.j} + \bar{y}_{..})^2$$

Since the product term will be vanishes:

$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = q\sum_i (y_{i.} - \bar{y}_{..})^2 + p\sum_j (y_{.j} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - y_{i.} - y_{.j} + \bar{y}_{..})^2$

TSS = SSA + SSB + SSE

Total Sum of Square = Sum of Square due to Factor A (Treatment) + Sum of Square due to Factor B (Blocks) + Sum of Square due to Error

Degrees of Freedom

TSS has n-1 degree of freedom
SSA has p-1 degrees of freedom
SSB has q-1 degrees of freedom
SSE has (p-1) (q-1) degrees of freedom.
Thus d.f. is also additive in nature.

Mean Sum of Square

Dividing the sum of squares by its degrees of freedom, we get corresponding variance or mean squares, therefore,

Mean Square due to factor A (MSA) = $\frac{SSA}{p-1}$

Mean Square due to factor B (MSB) = $\frac{SSB}{q-1}$

Mean Square due to error (MSE) = $\frac{SSE}{(p-1)(q-1)}$

F-test Statistics

The F-test statistics for factor A; $F_A = \frac{MSA}{MSE}$

If $F_A < F_{(p-1),\,(p-1)(q-1)}$, then our null hypothesis H_{0A} is true and we conclude that, there is no difference among factor A, otherwise reject.

The F-test statistics for factor B; $F_B = \frac{MSB}{MSE} \sim F_{(q-1),\,(p-1)(q-1)}$

If $F_B < F_{(q-1),\,(p-1)(q-1)}$, then our null hypothesis H_{0B} is true and we conclude that, there is no difference among factor B, otherwise reject.

4.5.4 ANOVA Table

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Sum of Squares*	*Variance Ratio*	
				$F_{cal.}$	$F_{tab.}$
Factor A	p-1	SSA	MSA= SSA/p-1	$F_A = \frac{MSA}{MSE}$	$F_{(p-1),\,(p-1)(q-1)}$
Factor B	q-1	SSB	MSB= SSB/q-1	$F_B = \frac{MSB}{MSE}$	$F_{(q-1),\,(p-1)\,(q-1)}$
Error	(p-1)*(q-1)	SSE	MSE= SSE/(p-1) (q-1)		
Total	n-1	TSS			

If $F_A < F_{(p-1),\,(p-1)(q-1)}$, then our null hypothesis H_{0A} is true and we conclude that, there is no difference among factor A, otherwise reject.

If $F_B < F_{(q-1),\,(p-1)(q-1)}$, then our null hypothesis H_{0B} is true and we conclude that, there is no difference among factor B, otherwise reject.

If F_A is significant, to find which pairs of treatment means differ significantly, we arrange mean yields in descending order of magnitude and then test for the significance of the pairwise differences by comparing them with critical difference:

$$C.D. = \sqrt{\frac{2*MSE}{r}} \text{ x } [t_{0.025}(for\ error\ d.f.)]$$

Where standard error (S.E.) of the treatment mean is given by: $S.E.(\bar{t_i}) = S_E/\sqrt{r}$; *(i= 1, 2,….t)*

4.5.5 Advantages and Disadvantages of RBD

Advantages

RBD has the several advantages which are given below:

1. This design is very flexible.
2. It is applicable to a moderate o. of treatments.
3. If we need extra replication for some treatment, there may be applied to be more than unit per block.
4. It also enables us to use different techniques to different block.

5. If any or all of the observation for particular block or treatment is missing, then we can estimate the values of the missing observation and perform the test.

Disadvantages

1. The main disadvantage of this design is that if the blocks are internally homogeneous then a large error term will amount.
2. If we increase the number of treatments, then block size is increase and in this case, we have lesser control over error.

Example: Three varieties A, B and C of a crop are tested in a randomized block design with four replications. The plot yields in pounds are as follows:

A (6)	C (5)	A (8)	B (9)
C (8)	A (4)	B (6)	C (9)
B (7)	B (6)	C (10)	A (6)

Analyse the experimental yield and state your conclusion.

Solution: Let us take the hypothesis that variation between varieties and between blocks do not differ significantly from the variance due to random error.

Calculation of Sum of Squares

Variety	***1***	***2***	***3***	***4***	***Total***
A	6	4	8	6	$V_1 = 24$
B	7	6	6	9	$V_2 = 28$
C	8	5	10	9	$V_3 = 32$
Total	$B_1 = 21$	$B_2 = 15$	$B_3 = 24$	$B_4 = 24$	$G = 84$

$$\text{Correction Factor } = \frac{G^2}{N} = \frac{(84)^2}{12} = 588$$

$$\text{Sum of Squares between Blocks (SSB) } = \frac{B_1^2+B_2^2+B_3^2+B_4^2}{3} - CF$$

$$= \frac{(21)^2+(15)^2+(24)^2+(24)^2}{3} - 588 = 18$$

$$\text{Sum of Squares between Varities (SSV) } = \frac{V_1^2+V_2^2+V_3^2}{4} - CF$$

$$= \frac{(24)^2+(28)^2+(32)^2}{4} - 588 = 8$$

$$\text{Total Sum of Squares (TSS) } = [6^2 + 4^2 + 8^2 + 6^2 + 7^2 + 6^2 + 6^2 + 9^2 + 8^2 + 5^2 + 10^2 + 9^2] - 588 = 36$$

Error Sum of Square (SSE) $= 36 - (18 + 8) = 10$

ANOVA Table

Source of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$F_{cal.}$	$F_{tab.}$
Between Varieties	3-1 = 2	8	$\frac{8}{2} = 4$	$\frac{4}{1.667} = 2.4$	$F_{2, 6, 0.05}=5.14$
Between Blocks	4-1 = 3	18	$\frac{18}{3} = 6$	$\frac{6}{1.667} = 3.6$	$F_{3, 6, 0.05}=4.76$
Error	(3-1)*(4-1) = 6	10	$\frac{10}{6} = 1.667$		
Total	12-1 = 11	36			

In both the cases, i.e., for variation between varieties and variation between blocks, the calculated value of F is less than the tabulated value, hence, the variances between varieties and between blocks do not differ significantly from the variance due to random error.

4.6 Latin Square Design (LSD)

In RBD, the principal of local control used by grouping the unit in one way, i.e., according to the block. Grouping can be carried in two ways each way corresponding to a source of variation among the units and get LSD. This design is frequently used in the agricultural field experiment where the fertility contour is not always known. Then the LSD eliminate the initial variability among the units into orthogonal direction. So, the LSD is also used in industry and laboratory.

In this design, the number of treatment equal to the common replication number per treatments as well as number of replications for each treatment. Then the total number of experimental units needed for this design will be m^2. These units are arranged in m rows and m columns, then the m treatments are allocated to these m^2 units at random subject to condition that each treatment occurs once and only once in each row and in each column.

The LSD is actually an Incomplete Three-Way Layout, where all the three factors, row, column and treatment are at the same number of levels (m). For three-way (complete) layout with each factor at m-level, we need m^3 experimental units, but in LSD, we use only m^2 units out of m^3 as following:

Let us consider a 4x4 Latin square for comparing four varieties of a crop. This case, we need 16 plots arranged in 4 rows and 4 columns. Let us represent the variety denoted by A, B, C and D. The Latin square will be:

Rows	*Column*			
	1	*2*	*3*	*4*
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

4.6.1 Layout

A n*n Latin square with n letters A, B, C, D, …….., in the natural order occurring in the first row and first column is called a standard square. For a standard n*n Latin square, we may obtain n!*(n-1)! Different LSD's by permuting all the n-columns and (n-1) row, except first row. Hence, there are in all n!*(n-1)! Different LSD with the same standard square.

4.6.2 Analysis

Let y_{ijk} be the observation on the treatment combination, where the factor A is at the i^{th} level, factor B is at the j^{th} level and factor C is at the k^{th} level.

For LSD, the linear model is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk};\ i, j, k = 0, 1, 2, \ldots\ldots\ldots\ldots m$$

Where, $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \tau_k = 0$ and e_{ijk} are assumed to be independently normal with mean 0 and variance σ_e^2.

α, β, τ denote the effect due to factor A or row, factor B or column and factor C or treatment respectively.

Here, the hypothesis of the interest is that all the treatment effects are zero; i.e.:

H_{oA}: There is no significant difference among rows.

H_{oB}: There is no significant difference among column.

H_{oT}: There is no significant difference among treatments.

Against,

H_{1A}: There is a significant difference among rows.

H_{1B}: There is a significant difference among column.

H_{1T}: There is a significant difference among treatments.

4.6.3 Least Square Estimates

To test the above hypothesis, we find the least square estimates by minimizing the residual sum of squares as:

$$S = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} e_{ijk}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}(y_{ijk} - \alpha_i - \beta_j - \tau_k)^2$$

The normal equations are as follows:

$\frac{dS.S.E}{d\mu} = 0; \frac{dS.S.E}{d\alpha} = 0; \frac{dS.S.E}{d\beta} = 0; \frac{dS.S.E}{d\tau} = 0$

We get the estimates as follows:

$$\hat{\mu} = \bar{y}_{...}$$

$$\hat{\alpha} = \bar{y}_{i..} - \bar{y}_{...}$$

$$\hat{\beta} = \bar{y}_{.j.} - \bar{y}_{...}$$

$$\hat{\tau} = \bar{y}_{..k} - \bar{y}_{...}$$

Now, substituting all these estimates in linear model of LSD, we have:

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..k} - \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})$$

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..k} - \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})$$

Now, squaring both sides and summing over all observations, we get:

$$\sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{...})^2 = \sum_i\sum_j\sum_k(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_i\sum_j\sum_k(\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_i\sum_j\sum_k(\bar{y}_{..k} - \bar{y}_{...})^2 - \sum_i\sum_j\sum_k(\bar{y}_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2$$

Since the product term will vanish, hence:

$$\sum_i\sum_j\sum_k\left(y_{ijk} - \bar{y}_{...}\right)^2 = m^2\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2 + m^2\sum_j(\bar{y}_{.k.} - \bar{y}_{...})^2 + m^2\sum_k(\bar{y}_{..k} - \bar{y}_{...})^2 - \sum_i\sum_j\sum_k\left(\bar{y}_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...}\right)^2$$

$$TSS = SSR + SSC + SST + SSE$$

Where,

Total Sum of Square (TSS) $= \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{...})^2$

Sum of Square due to Factor A or Row (SSR) $= m^2\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$

Sum of Square due to Factor B or Column (SSC) $= m^2 \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$

Sum of Square due to Factor C or Treatment (SST) $= m^2 \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2$

Sum of Square due to Error (SSE) $= \sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2$

Degree of Freedom

TSS has m^2-1 degrees of freedom

SSR has m-1 degrees of freedom

SSC has m-1 degrees of freedom

SST has m-1 degrees of freedom

SSE has (m-1)*(m-2) degrees of freedom

In this way, we see that the degree of freedom is additive in nature.

Mean Sum of Squares

Mean Sum of Square due to Factor A or Row (MSR) = $\frac{SSR}{m-1}$

Mean Sum of Square due to Factor B or Column (MSC) = $\frac{SSC}{m-1}$

Mean Sum of Square due to Factor C or Treatment (MST) = $\frac{SST}{m-1}$

Mean Sum of Square due to Error (MSE) = $\frac{SSE}{(\mathrm{m}-1)*(\mathrm{m}-2)}$

F-test Statistics

F-test statistic for Factor A or Row: $F_A = \frac{MSR}{MSE} \sim F_{(m-1),(m-1)(m-2),\alpha}$

If $F_A > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oA} is rejected.

F-test statistic for Factor B or Column: $F_B = \frac{MSC}{MSE} \sim F_{(m-1),(m-1)(m-2),\alpha}$

If $F_B > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oB} is rejected.

F-test statistic for Factor C or Treatment: $F_T = \frac{MST}{MSE} \sim F_{(m-1),(m-1)(m-2),\alpha}$

If $F_T > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oT} is rejected.

4.6.4 ANOVA Table

Source of Variation	*Degree of Freedom*	*Sum of Square*	*Mean Sum of Square*	*Variance Ratio*	
				***F**cal.*	***F**tab.*
Row	m-1	SSR	MSR	F_A	$F_{(m-1),(m-1)(m-2),\alpha}$
Column	m-1	SSC	MSC	F_B	
Treatment	m-1	SST	MST	F_T	
Error	(m-1)*(m-2)	SSE	MSE		
Total	$m^2 - 1$	TSS			

If $F_A > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oA} is rejected.

If $F_B > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oB} is rejected.

If $F_T > F_{(m-1),(m-1)(m-2),\alpha}$, then our null hypothesis H_{oT} is rejected.

Standard error (S.E.) of the difference between any two-treatment means is given by:

S.E., $(\bar{t_i} - \bar{t_j}) = \sqrt{(s_E^2\,(\frac{1}{m}+\frac{1}{m}))} = \sqrt{(2s_E^2/m)}$

And the critical difference (C.D.) for the significance of the difference between any two treatment means at level of significance (α) is given by:

C.D. $(\bar{t_i} - \bar{t_j}) = t_{(error\ d.f.)}\left(\frac{\alpha}{2}\right) x$ S.E., $(\bar{t_i} - \bar{t_j}) \quad = \quad t_{(m-1)(m-2)}(0.025)\ x \sqrt{(2s_E^2/m)}$

4.6.4 Advantages and Disadvantages of LSD

Advantages

If we don't have direction of fertility gradient, LSD provides to eliminate from the error, two major sources of the variation that are not relevant to the comparison to y grouping the units in two phases. Thus, LSD is an improvement over RBD in controlling the error. Since LSD is an incomplete three-way layout, therefore we needed only m^2 observations out of m^3.

Generally, in the field experiment, the plots are laid out in a square, but LSD maybe used with the plots in continuous. *Example*: the fertility gradient in all along the line.

Disadvantages

In LSD, the number of replications must be same as the number of treatment and hence, the square larger than 12x12 are seldom used, because the size of square becomes large, resulting square does not remain homogenous, on the other hand, a small square provide only a few degrees of freedom for error. Therefore, the standardized size for LSD, which are commonly used are 5x5 and 8x8.

4.7 Estimation of Missing Values in Randomized Block Design (RBD)

There are some following ways :

4.7.1 One missing value in RBD

Let the observation y_{ij} = x (say), in the j^{th} block and receiving i^{th} treatment be missing as shown in the following table:

Treatment	***Blocks***						***Total***
	B_1	***B_2***	**………...**	***B_j***	**………...**	***B_q***	
1	y_{11}	y_{12}	**………...**	y_{1j}	………...	y_{1q}	$T_{1.}$
2	y_{21}	y_{22}	**………...**	y_{2j}	**………...**	y_{2q}	$T_{2.}$
. . .							. . .
i	y_{i1}	y_{i2}	**………...**	Missing (x)	**………...**	y_{iq}	$T'_{i.}$+x
. . .							. . .
p	y_{k1}	y_{k2}	**………...**	y_{pj}	**………...**	y_{pq}	$T_{p.}$
Total	$T_{.1}$	$T_{.2}$	**………...**	$T'_{.j}$+x	**………...**	$T_{.q}$	$T_{..}=T'_{..}+x$

Where,

$T'_{i.}$ – Total of known observation getting the i^{th} treatment

$T'_{.j}$ - Total of known observation getting the j^{th} treatment

$T'_{..}$ - Total of all known observations

Now, we have:

$$\text{RSS} = \sum_i \sum_j {y_{ij}}^2 + x^2$$

$$\text{CF} = \frac{(T'_{..}+x)^2}{n} = \frac{(T'_{..}+x)^2}{pq}$$

TSS = RSS – CF

$$\text{TSS} = x^2 + \text{constant w.r.t. } x - \frac{(T'_{..}+x)^2}{n}$$

$\text{SST} = \frac{1}{q}[(T'_{i.} + x)^2 + \text{constant w.r.t. x}] - \text{CF}$

$\text{SSB} = \frac{1}{p}[(T'_{.j} + x)^2 + \text{constant w.r.t. x}] - \text{CF}$

$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB}$

$$\text{SSE} = x^2 + \text{constant w.r.t. x} - \text{CF} - \left[\frac{1}{q}[(T'_{i.} + x)^2 + \text{constant w.r.t. x}] - \text{CF}\right] - \left[\frac{1}{p}[(T'_{.j} + x)^2 + \text{constant w.r.t. x}] - \text{CF}\right]$$

$$\text{SSE} = x^2 - \frac{1}{q}[(T'_{i.} + x)^2] - \frac{1}{p}[(T'_{.j} + x)^2] + \frac{(T'_{..} + x)^2}{n} + \text{constant w.r.t. x}$$

To estimate x, we minimize the error. Thus, minimize the value of SSE we have:

$$\frac{dS.S.E}{dx} = 0 \Rightarrow 2x - \frac{2}{q}[(T'_{i.} + x)] - \frac{2}{p}[(T'_{.j} + x)] + \frac{2}{n}[(T'_{..} + x)] = 0$$

$pkx - pT'_{i.} - qT'_{.j} - px - qx + x + T'_{..} = 0$

$x(pk - p - k + 1) = pT'_{i.} + qT'_{.j} - T'_{..}$

$$\hat{x} = \frac{p * T'_{i.} + q * T'_{.j} - T'_{..}}{(p-1) * (q-1)}$$

4.7.2 Two Missing values in RBD

Let two missing observations be x and y as shown in the following table:

Treatment	***Blocks***					***Total***
	B_1	……...	B_j	……...	B_q	
1	y_{11}	……...	Missing (x)	……...	y_{1q}	$R_1 + x$
. . .		……...		……...		. . .
i	y_{i1}	……...	yij	……...	y_{iq}	R_i
. . .		……...		……...		. . .
p	y_{k1}	……...	y_{pj}	……...	Missing (y)	$R_2 + y$
Total	$T_{.1}$	……...	$C_1 + x$	……...	$C_2 + y$	$T_{..} = T'_{..}+x$

Let R_1 and R_2 be the total of known observations in the row containing x and y respectively and C_1 and C_2 be the total of known observations in the column containing x and y respectively. Let S be the total of known observations as shown in the above table.

Now, we have:

$$RSS = \sum_i \sum_j {y_{ij}}^2 + x^2 + y^2$$

$$CF = \frac{(S+x+y)^2}{n} = \frac{(S+x+y)^2}{pq}$$

TSS = RSS – CF

$$TSS = x^2 + y^2 + constant\ w.r.t.x - \frac{(S+x+y)^2}{n}$$

$$SST = \frac{1}{q}[(R_1+x)^2 + (R_2+x)^2] + constant\ w.r.t.x - \frac{(S+x+y)^2}{n}$$

$$SSB = \frac{1}{p}[(C_1+x)^2 + (C_2+x)^2] + constant\ w.r.t.x - \frac{(S+x+y)^2}{n}$$

$$SSE = TSS - SST - SSB$$

$$SSE = x^2 + y^2 + constant\ w.r.t.x - \frac{(S+x+y)^2}{n} - \frac{1}{q}[(R_1+x)^2 + (R_2+x)^2] -$$

$$\frac{1}{p}[(C_1+x)^2 + (C_2+x)^2]$$

To estimate x and y we minimize the SSE. Thus, for minimizing the value of SSE, we have:

$\frac{dS.S.E}{dx} = 0$ and $\frac{dS.S.E}{dy} = 0$

Now,

$\frac{dS.S.E}{dx} = 0 \Rightarrow 2x - \frac{2}{q}[(R_1+x)] - \frac{2}{p}[(C_1+x)] + \frac{2}{pq}[(S+x+y)] = 0$

$pqx – p(R_1+x) – q(C_1+x) + S+x+y = 0$

$(pq – p – q + 1)x – pR_1 – qC_1 + S + y = 0$

$(p-1)*(q-1)*x = pR_1 +qC_1 – S – y = 0$

$$\hat{x} = \frac{p*R_1 + q*C_1 - S - y}{(p-1) * (q-1)}$$

$\frac{dS.S.E}{dy} = 0 \Rightarrow 2y - \frac{2}{q}[(R_2+y)] - \frac{2}{p}[(C_2+y)] + \frac{2}{pq}[(S+x+y)] = 0$

$pqy – p(R_2+x) – q(C_2+y) + S + x + y = 0$

$(pq – p – q + 1)y – pR_1 – qC_1 + S + y = 0$

$(p-1)(q-1)*y = pR_2 +qC_2 – S – y = 0$

$$\hat{y} = \frac{p * R_2 + q * C_2 - S - x}{(p-1) * (q-1)}$$

Solving these two equations, we get the estimations of x and y.

Statistical Analysis

Analysis of variance is performed in the usual way after substituting the estimated value of missing observations. For each missing observations, one degree of freedom is subtracted from total and also from error degree of freedom. The adjusted treatment sum of squares is obtained by subtracting the Adjustment Factor from the SST, where adjusted factor is given as:

$$AF = \frac{[pT'_{i.} + T'_{.j} - T'_{..}]^2}{p(p-1)(q-1)^2}$$

Example: The yields of 6 varieties in a 4-replicate experiment for which one value is missing are given below. Estimate the missing value and analyse the data.

Blocks	***Treatments***					
	1	***2***	***3***	***4***	***5***	***6***
1	18.5	15.7	16.2	14.1	13.0	13.6
2	11.7	-	12.9	14.4	16.9	12.5
3	15.4	16.6	15.5	20.3	18.4	21.5
4	16.5	18.6	12.7	15.7	16.5	18.0

Solution: For estimating one missing value in a RBD, the formula is

$$\hat{x} = \frac{p * T'_{i.} + q * T'_{.j} - T'_{..}}{(p-1) * (q-1)}$$

Blocks	***Treatments***						***Block Totals (Bj)***
	1	***2***	***3***	***4***	***5***	***6***	
1	18.5	15.7	16.2	14.1	13.0	13.6	91.1
2	11.7	-	12.9	14.4	16.9	12.5	68.4
3	15.4	16.6	15.5	20.3	18.4	21.5	107.8
4	16.5	18.6	12.7	15.7	16.5	18.0	98.0
Treatment Totals (Ti)	62.1	50.9	57.3	64.5	64.8	65.7	365.3

From the above data, we have:

p = 6, $T'_{i.}$ = 50.9, q = 4, $T'_{.j}$ = 68.4, $T'_{..}$ = 365.3

$$\hat{x} = \frac{(6 * 50.9) + (4 * 68.4) - 365.3}{(6 - 1) * (4 - 1)} = \frac{213.7}{15}$$

$$\hat{x} = 14.25$$

The null hypotheses for testing are:

H_{oT}: Treatments are homogenous; H_{oB}: Blocks are homogenous

Against,

H_{1T}: At least two treatments are different; H_{1B}: At least two blocks are different

Substituting the value of $\hat{x}$ is the table, we have:

Treatment Totals:	T_1	T_2	T_3	T_4	T_5	T_6
	62.1	65.15	57.3	64.5	64.8	65.7
Block Totals:	B_1	B_2	B_3	B_4		
	91.1	82.65	107.8	98.0		

G = 379.55, N = 4*6 = 24

$$\text{Correction Factor (CF)} = \frac{(379.55)^2}{24} = 6002.42$$

$$\text{Residual Sum of Square (RSS)} = \sum_{i=1}^{6}\sum_{j=1}^{4} T_{ij}^2 = 6151.94$$

$$\text{Total Sum of Square (TSS)} = \text{RSS} - \text{CF} = 6151.94 - 6002.42 = 149.52$$

$$\text{Sum of Square due to Treatments (SST)} = \frac{1}{4}\sum_{i=1}^{6} T_i^2 = \frac{24060.0025}{4} - 6002.42 = 12.58$$

$$\text{Sum of Square due to Blocks (SSB)} = \frac{1}{6}\sum_{j=1}^{4} B_j^2 = \frac{36355.07}{6} - 6002.42 = 56.76$$

$$\text{Sum of Square due to Error (SSE)} = \text{TSS} - \text{SST} - \text{SSB} = 80.18$$

The adjustment factor for treatment sum of square is given by:

$$AF = \frac{\left[pT'_{i.} + T'_{.j} - T'_{..}\right]^2}{p(p-1)(q-1)^2} = \frac{[(6*50.9) + 68.4 - 365.3]^2}{6(6-1)(4-1)^2} = 0.267$$

Therefore, adjusted value of Sum of Square due to Treatment = 12.580 – 0.267 = 12.313

ANOVA Table

Source of Variation	*Degree of Freedom*	*Sum of Square*	*Mean Sum of Square*	*Variance Ratio*	
				$F_{Cal.}$	$F_{Tab.}$
Treatments (Adjusted)	*5*	*12.313*	$\frac{12.313}{5} = 2.06$	$\frac{2.06}{5.72} = 2.33$	$F_{0.05(5,14)} = 2.96$
Blocks	*3*	*56.760*	$\frac{56.760}{3} = 18.92$	$\frac{18.92}{5.72} = 3.30$	$F_{0.05(3,14)} = 3.34$
Error	*14**	*80.45*	$\frac{80.45}{14} = 5.72$		
Total	*22**				

** Here, one degree of freedom is lost for total sum of square and consequently for error sum of square due to the estimation of missing value from the given data.*

From the ANOVA table, we see that calculated F values of both treatments and blocks are insignificant, and consequently, H_{ob} and H_{oT} maybe retained, i.e., we may regard the treatments as well as blocks to be homogenous.

4.8 Estimation of Missing Values in Latin Square Design (LSD)

Let the observation y_{ijk} = x (say) be the missing value in the i^{th} row and j^{th} column and receiving k^{th} treatment in m*m Latin square. Let:

R – Total of known observations in the i^{th} row

C – Total of known observations in the j^{th} column

T – Total of known observations in the k^{th} treatment

S – Total of known observations

Now, we have:

$RSS = \sum_i \sum_j {y_{ij}}^2 + x^2$

$CF = \frac{(S+x)^2}{m^2}$

TSS = RSS – CF

$TSS = x^2 + constant\ w.r.t.x - \frac{(S+x)^2}{m^2}$

$SSR = \frac{1}{m}[(R+x)^2] + constant\ w.r.t.x - \frac{(S+x)^2}{m^2}$

$SSC = \frac{1}{m}[(C+x)^2] + constant\ w.r.t.x - \frac{(S+x)^2}{m^2}$

$SST = \frac{1}{m}[(T+x)^2] + constant\ w.r.t.x - \frac{(S+x)^2}{m^2}$

$SSE = TSS - SSR - SSC - SST$

$SSE = x^2 - \frac{1}{m}[(R+x)^2] - \frac{1}{m}[(C+x)^2] - \frac{1}{m}[(T+x)^2] + constant\ w.r.t.x + \frac{2(S+x)^2}{m^2}$

Now, to estimate x we minimize SSE. Thus, for minimum value of SSE:

$\frac{dS.S.E}{dx} = 0 \Rightarrow 2x - \frac{2}{m}[(R+x)] - \frac{2}{m}[(C+x)] - \frac{2}{m}[(T+x)] + \frac{4}{m^2}[(S+x)] = 0$

$m^2x - Rm - mx - Cm - mx - Tm - mx + 2S + 2x = 0$

$(m^2 - 3m + 2)*x - m*(R + C + T) + 2S = 0$

$(m - 1)*(m - 2)*x = m*(R + C + T) - 2S$

$$\hat{x} = \frac{m(R+C+T) - 2S}{(m-1) * (m-2)}$$

Statistical Analysis

Analysis of Variance is performed in the usual way after substituting the estimated value of missing observations. For each missing observations, one degree of freedom is subtracted from total and also from error degree of freedom. The adjusted treatment sum of squares is obtained by subtracting the Adjustment Factor from the SST, where adjusted factor is given as:

$$AF = \frac{[(m-1)T + R + C - S]^2}{[(m-1)(m-2)]^2}$$

Example: An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used were as follows:

A: No clay

B: Clay at 100 per acre

C: Clay at 200 per acre

D: Clay at 300 per acre.

The yields were in plots of 8 metres and are given below:

Rows	***Columns***			
	I	***II***	***III***	***IV***
I	D 29.1	B 18.9	C 29.4	A 5.7
II	C	A	D	B

	16.4	10.2	21.2	19.1
III	A 5.4	D 38.8	B 24.0	C 37.0
IV	B 24.9	C 41.7	A 9.5	D 28.9

a) Perform the ANOVA and calculate the critical difference for the treatment mean yields.

b) Yield under 'A' in the first column was missing. Estimate the missing value and carry out the ANOVA.

Solution-

Rows	***Columns***				***Row total (R_i)***
	I	***II***	***III***	***IV***	
I	D 29.1	B 18.9	C 29.4	A 5.7	83.1
II	C 16.4	A 10.2	D 21.2	B 19.1	66.9
III	A 5.4	D 38.8	B 24.0	C 37.0	105.2
IV	B 24.9	C 41.7	A 9.5	D 28.9	105.0
Column Total (C_j)	75.8	109.6	84.1	90.7	360.2

The four treatment totals are: A: 30.8, B: 86.9, C: 124.5, D: 118.0

Grand total (G) = 360.2, N= 16

Correction factor (C.F.) = $\frac{G^2}{N}$ $\frac{(360.2)^2}{16}$ = 8109.0025

Raw S.S. = $\sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}2$ = $(29.1)^2+(18.9)^2$ + ……….+ $(9.5)^2 + (28.9)^2$ = 10,052.08

Total S.S. = 10,052.08-8,09.0025 = 1,943.0775

S.S.R. = $\frac{1}{4}$ $[(83.1)^2+(66.9)^2 + (105.2)^2 + (105.0)^2]$ – 8,109.0025

= $\frac{33,473.26}{4}$ – 8,109.0025 = 259.3125

S.S.C. = $\frac{1}{4}$ $[(75.8)^2+(109.6)^2 + (84.1)^2 + (90.7)^2]$ – 8,109.0025

= $\frac{33057.10}{4}$ – 8,109.0025 = 155.2725

S.S.T. = $\frac{1}{4}$ $[(30.8)^2+(86.9)^2 + (124.5)^2 + (118.0)^2]$ – 8,109.0025

= $\frac{37924.50}{4}$ – 8,109.0025 = 1372.1225

Error S.S. = TSS-SSR-SSC-SST = 156.3700

ANOVA Table for LSD

Source of Variation (1)	***Degree of Freedom (2)***	***Sum of Square (3)***	***Mean Sum of Square (4) = (3) / (2)***	***Variance Ratio***	
				F_{Cal}.	***F_{Tab}.***
Rows	3	259.5375	86.4375	3.32	$F_{3,6}$ (0.05) = 4.76
Columns	3	155.2725	51.7575	1.98	
Treatments	3	1,372.1225	457.3742	17.55	
Error	6	156.3700	26.0616		
Total	15	1,943.0775			

Hence, we conclude that the variation due to rows and columns is not significant but the treatments, i.e., different levels of clay, have significant effect on the yield. To determine which of the treatment pairs differ significantly, we have to calculate the critical difference (C.D.)

S.E. of difference between any two treatment means= $S_E = \sqrt{\left(\frac{2}{m}\right)} = \sqrt{\left(2 * \frac{26.0616}{4}\right)}$ = 3.609

Hence C.D. = 3.609 * $t_{0.025}$ (for error d.f.) = 3.609 * 2.447= 8.83

We now arrange treatment means in their decreasing order of magnitude as given as follows:

Treatment	**Mean Yield**	$\|\Delta = \boldsymbol{Difference}\|$
C	31.1250	1.625
D	29.5000	7.775
B	21.7260	14.025
A	7.7000	

We, therefore, conclude that:

i. The difference between mean yields of C and D is not significant and they may therefore be regarded alike as regards their effect on yield. Similarly, argument holds for the pair D and B.

ii. The treatment C and B are significant from each other as regards their effect on yields, since the difference between their mean yields, viz., 9.4 exceeds the C.D. As such treatment C is to treatment B. Similar argument holds for any other pair left.

(b) The formula for obtaining the missing value in LSD is: $\hat{x} = \frac{m(R+C+T)-2S}{(m-1)*(m-2)}$

$\hat{x} = \frac{4\,(99.8+70.4+25.4)-2*354.8}{3*2} = 12.13$

As a result of replacing the missing figure by its estimate 12.13, *Corrected or Adjusted S.S.* are obtained as follows:

Raw SS = 10,052.08- $(5.4)^2 + (12.13)^2$ = 10,170.06

$G^2 = (360.2 - 5.4 + 12.13)^2 = (366.93)^2 = 134637.62$

C.F. = $\frac{G^2}{N}\ \frac{134637.62}{16} = 8414.85$

TSS = RSS-CF= 10170.06-8414.85= 1755.21

$\sum R_i^2 = 33473.26 - (105.2)^2 + (105.2 - 5.4 + 12.13)^2$ = 34,934.54

SSR= $\frac{34,934.54}{4} - 8,414.85 = 318.79$

$\sum C_j^2 = 33,057.10 - (75.8)^2 + (75.8 - 5.4 + 12.13)^2$ = 34,122.66

SSC= $\frac{34122.66}{4} - 8,414.85 = 115.81$

$\sum T_k^2 = 37,924.50 - (30.8)^2 + (30.8 - 5.4 + 12.13)^2$ =38,384.36

Adjusted factor for treatment SS is: $AF = \frac{[(m-1)T + R + C - S]^2}{[(m-1)(m-2)]^2}$

$AF = \frac{(3*25.4+99.8+70.4-354.8)^2}{(3*2)^2}$ = 326.40

Adjusted SST = $\frac{38,384.36}{4}$ −8,414.85-326.40= 854.84

Adjusted SSE = TSS-SSR-SSC-SST(Adjusted)

= 1,755.21-318.79-115.81-854.84 = 465.77

Corrected ANOVA for LSD (Missing Observation)

Source of Variation ***(1)***	***Degree of Freedom*** ***(2)***	***Sum of Square*** ***(3)***	***Mean Sum of Square*** ***(4) = (3)/ (2)***	***Variance Ratio***
Rows	3	318.79	106.26	1.14
Columns	3	115.81	38.60	<1
Treatments (adjusted)	3	854.84	284.95	3.06
Error (adjusted)	14-9 = 5*	465.77	93.15	
Total	15-1 = 14*	1,455.21		

***** 1 d.f. is reduced for the d.f. of total SS and consequently for Error SS because one missing observation has been estimated and its estimated value is used in computing the various SS.

Tabulated $F_{0.05}(3,5) = 5.41$

Since the calculated values of F_R, F_C and F_T are less than the tabulated value, more of them is significant. Hence, the treatments do not differ significantly.

4.9 Self-Assessment Exercise

Question-1: An experiment was carried out in a pharmaceutical firm to find out whether there were any differences in the disintegration time of five different types of caplets. The following data were obtained. Analyse the given data.

Type of Caplet (i)	*Disintegration time for n_i caplets (seconds)*
1	2,6,4,8,6,7 ($n_1 = 6$)
2	3,7,6,4,8 ($n_1 = 5$)
3	4,8,10,7,9,11 ($n_1 = 6$)
4	10,12,9,7,8 ($n_1 = 6$)
5	12,7,9,8,11,13,9 ($n_1 = 6$)

Question-2: Three different washing solutions are being compared to study their effectiveness in retarding bacteria growth in five-gallon milk containers. The analysis is done in a laboratory, and only three trails can be run on any day. Because days could represent a potential source of variability, the experimenter decides to use a randomized block design. Observations are taken for four days, and the data are shown here. Analyse the data and draw conclusions.

Solution	*Days*			
	1	*2*	*3*	*4*
1	13	22	18	39

2	16	24	17	44
3	5	4	1	22

Question-3: A problem was posed to estimate the petrol consumption rates of the four 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 of 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 ran with different speeds on different days. The speeds were 25, 35, 50, 60 and 70 mph. Which car was to be 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 4 Makes of cars. For each car the number of miles covered per gallon of petrol was observed. The observations are presented below. Analyse the given randomised block design and interpret the results.

Miles per Gallon of Petrol

Makes of car	*Speeds of the cars in miles per hour (mph)*				
	25	*35*	*50*	*60*	*70*
A	20.6	19.5	18.1	17.9	16.0
B	19.5	19.0	15.6	16.7	14.1
C	20.5	18.5	16.3	15.2	13.7
D	16.2	16.5	15.7	14.8	12.7

Question-4: To determine the petrol consumption rates of the four different makes of cars, five drivers were chosen and each driver was used on one of 5 days. On that day he drove five cars each of a different make and each car with a different speed. The arrangement of drivers, speeds and makes was in the given table. Analyse the given block design and interpret the results.

Drivers and Days	*Speeds in miles per hour*				
	25	*35*	*50*	*60*	*70*
D_1	B (19.5)	E (21.7)	A (18.1)	D (14.8)	C (13.7)
D_2	D (16.2)	B (19.0)	C (16.3)	A (17.9)	E (17.5)
D_3	A (20.6)	D (16.5)	E (19.5)	C (15.2)	B (14.1)
D_4	E (22.5)	C (18.5)	D (15.7)	B (16.7)	A (16.0)
D_6	C (20.5)	A (19.5)	B (15.6)	E (18.7)	D (12.7)

Question-5: To compare the effects of six treatments, an experiment with these treatments was conducted in four randomised blocks. One observation under treatment 1 in block 1 was lost accidentally. Available observations along with the layout are given below. Treatments are indicated by numbers within parentheses. Analyse the data and draw conclusions.

Block	*Treatment and Yield (in certain units)*					
	1	*2*	*3*	*4*	*5*	*6*
1	(1) -	(3) 27.7	(2) 20.6	(4) 16.2	(5) 16.2	(6) 24.9
2	(3) 22.7	(2) 28.8	(1) 27.3	(4) 15.0	(6) 22.5	(5) 17.0
3	(6) 26.3	(4) 19.6	(1) 38.5	(3) 39.5	(2) 39.5	(5) 15.4
4	(5) 17.7	(2) 31.0	(1) 28.5	(4) 14.1	(3) 34.9	(6) 22.6

4.10 Summary

This unit gives a complete knowledge about the terminology and basic principles of design of experiments, Complete Randomized Designs (CRD), Randomized Block Designs (RBD) and Latin Square Designs (LSD), analysis with missing observations for both RBD and LSD.

4.11 References

- Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer-Verlag, New York.
- Gupta S.C., Kapoor V.K. (2007). *Fundamentals of Applied Statistics*, Sultan Chand & Sons.
- Gupta S.P. (2021). *Statistical Methods*, Sultan Chand & Sons.
- Gupta V.K. & Nigam A.K. Nigam (1978-79). *Handbook an analysis of Agriculture Experiment,* IASRI Publication.

4.12 Further Reading

- Cochran, W. G. and Cox, G. M. (1957). *Experimental Design*. John Wiley & Sons, New York.
- Cochran, W.G. and Cox, G.M. (2003). *Experimental Design*, Asia Publishing House
- Das, M. N. and Giri, N. S. (1986). *Design and Analysis of Experiments (2nd Edition)*. Wiley.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*. New Age International (P) Ltd. New Delhi.
- Montgomery, D.C. (2017). *Design and Analysis of Experiments (9th Edition)*. John Wiley & Sons.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments,* Wiley, New York.

UNIT:5 FACTORIAL EXPERIMENTS

Structure

5.1 Introduction

A factorial experiment involves simultaneously more than one factor each at two or more levels. If the number of levels of each factor in an experiment is the same, then the experiment is called Symmetric Experiment; otherwise, it is called Asymmetric Experiment or Mixed Factorial Experiment.

These experiments provide an opportunity to study the individual effects of each factor and their interaction effect. When experiment is conducted factor by factor changing the level of a factor at time and keeping the other factors at constant levels, the interaction effects cannot be investigated.

Example: In many biological and clinical trials, factors are likely to have interaction. Therefore, factorial experiments are more informative in such investigations. The other advantages of these experiment are that these are helpful in economizing on experimental resources. When an experiment is conducted factor by factor, then much more resources are required for the same precision than what they are required in factorial experiment.

5.2 Objectives

After going through this unit, you should be able to:

- Know and understand benefits of using factorial experiments
- Understand the concept of 2^n factorial experiments, focusing on 2^2 and 2^3 factorial experiments.
- Understand the concept of 3^n factorial experiments, focusing on 3^2 and 3^3 factorial experiments.
- Applying the Yate's Method and analysis of variance test for 2^n and 3^n factorial experiments.

5.3 Advantages of Factorial Experiments

The factorial experiments have many advantages, few of them are:

1. It increases the scope of the experiment and gives information not only on the main factors but also on their interaction.
2. The various levels of one factor constitute replications of other factors and increase the information obtained on all factors.
3. When there are no interactions, factorial gives the maximum efficiency in estimating the given effects.
4. When interaction exists, their nature being unknown a factorial design to avoid misleading conclusions
5. In the factorial experiment the effects of a factor is estimated at several levels of other factors and conclusion hold over a wide range of conditions.

5.4 2^n Factorial Experiments

2^n factorial designs are the simplest class of factorial design involving factors at two levels, n being the no. of factors. The two levels can be denoted as lower level and higher level. The factors are usually divided by the capital letters and the two levels of the factor by, lower level with (1) and the corresponding small title letters. The treatment combinations can used two the individual of the symbols used for the individuals' levels. Thus, with three factors, A, B, and C the treatment combinations with all the factors at the higher level are written as abc. With two factors A and B the four treatment combinations for 2^2 factorial experiments are:

(1) → A and B at lower level

a → A at higher level and B at lower level

b → A at lower level and B at higher level.

ab → Both A and B at higher level

Standard order of Treatment Combinations

The complete list of 2^n treatment combinations can be conveniently written in standard order as follows:

Factorial Design	***Factors***	***Order***
2^1	A	(1), a
2^2	A and B	(1), a, b, ab
2^3	A, B and C	(1), a, b, c, ab, ac, bc, abc
2^4	A, B, C and D	(1), a, b, c, d, ab, ac, ad, bc, bd, cd, abc, abd, acd, bcd, abcd

Main effects and interaction effects

The capital letters A and B also serve to represent the main effects and interaction. The first order (two factors) interactions are denoted by AB, AC, BC, etc.; the 2^{nd} order (three-factor) interactions by ABC, ABD, BCD etc.

In a 2^n experiments, each main effect and interaction has one def. there are n main effects. ${}^nC_1$ first order interaction, ${}^nC_2$ 2^{nd} order and so on -------. The effect of A at the first level of B is (a) – (1). Similarly, the effect of A at the 2^{nd} level of B is (ab)-(b).

These two effects are called simple effects of factor A. The average effect of A over the two levels of B.

$$\text{A} = \frac{1}{2}[(ab) - (b) + (a) - (1)] = \frac{1}{2}[(a-1)(b+1)]$$

Similarly, the effect of B at the first level of A is (b)-(1) and the second level of A is (ab)-(a)

$$\text{B} = \frac{1}{2}[(ab) - (a) + (b) - (1)] = \frac{1}{2}[(a+1)(b-1)]$$

The interaction effect AB is $\text{AB} = \frac{1}{2}(a-1)(b-1)$

The main effects A, B and AB are contracts of treatment means for

$$\text{A} = \frac{1}{2}[(ab) - (b) + (a) - (1)] = \sum_{i=1}^{4} c_i t_i$$

Where,

$$C_1 = \frac{1}{2}, \quad C_2 = \frac{1}{2}, \quad C_3 = \frac{1}{2}\ , \quad C_4 = \frac{1}{2}$$

$$t_1 = (ab), \quad t_2 = (b), \quad t_3 = (a) \quad t_4 = (1)$$

$$\sum_i C_i = 0$$

5.4.1 2^2 Factorial Experiment

Suppose we have two factors A and B each of two levels (a, b). The first level of factor A and B generally expressed by the absence of the corresponding letter in the treatment combination. the 4-treatment combination are as follows:

1 → Factor A and B both are first level

a → Factor A at second and factor B beat first level

b → factor B at the second level and factor A at first level

ab → Both factors A and b are at second level

Notes: These 4 above treatments combination may be compare CRD, RBD and LSD. for a 2^2 experiment with q randomized blocks, the analysis will be same as the number of treatment combination p=4 in RBD and the analysis of 2^2 experiment in LSD with m=4.

Main Effect and Interaction Effect

The symbols [a] and [b] will be used to denote 4 means respectively of all the observations receiving the treatment combination a. the letters A, B and AB are used to denote for main effect due to factor A and B and the interaction between A and B. Consider the effect of A, the effect of changing the A from its first level to a 2nd in the presence of first level of factor B is given by (a)-(1) and the effect of changing factor A from its first level to second level in the presence of second level of B is given by (ab)-(b). these two effects are known as the simple effects of factor A. if A and B are independent in their effects then we assume that above two simple effects are equal and the average of these two simple effects is called the main effect due to factors A. Thus, main effect of factor A is given by

$$\mathrm{A} = \frac{1}{2}(a) - (1) + (\mathrm{ab}) - (\mathrm{b}) = \frac{1}{2}(a-1)(b+1)$$

Similarly, the main effect of factor B is given by:

$$\mathrm{B} = \frac{1}{4}(a+1)(b-1)$$

Note: Above two combinations are contrast and orthogonal too.

The main effect A is a linear function of the four treatment means with sum of the coefficient of the linear function=0.

Therefore, this is a contrast of treatment if the two factor are not independent then simple effect of A will not be the same and measure of independence and interaction of two factor A and B is defined as the ½ of the difference of the first simple effect i.e. {(a)-(1)} from the second simple effect i.e. {(ab)-(b)}. Thus, the interaction effect AB is given by:

$$\mathrm{AB} = \frac{1}{2}[(ab) - (b) - (\mathrm{a}) + (1)]$$

$$\mathrm{AB} = \frac{1}{2}(\mathrm{a} - 1)(\mathrm{b} - 1)$$

This is also a contrast of treatment means if we take the contrast A and AB, then it can be easily seen that the sum of the product of coefficient of these contrast A and AB is zero. it means that these two contrasts are orthogonal contrast. If we take the contrast B and AB, then it can be easily seen that the sum of the product of the coefficient of these contrast B and Ab is zero. it means that these two contrasts are orthogonal contrast.

Table for sign and divisors given general mean, main effect A, B and interaction Ab in terms of treatment mean is given as:

Effect	***Treatment means***				***Divisors***
	(1)	***(a)***	***(b)***	***(ab)***	
M	+	+	+	+	4
A	-	+	-	+	2
B	-	-	+	+	2
AB	+	-	-	+	2

Analysis of Variance for a 2^2-factorial experiment in q randomized blocks

The sum of squares due to the factorial effect can be obtained by multiplying the squares of the factorial totals by a suitable quantity. Here each sum of squares a single degree of freedom and sum of squares due to treatments have 3 degrees of freedom. Here it is to note that the factorial effect and their sum of squares from the treatment totals are easier to obtain rather than the treatment means. The effect totals can be written as

[A] = [ab] + [a] - [b] - [1]

[B] = [ab] + [b] - [a] - [1]

[AB] = [ab] – [b] - [a] + [1]

Since we use the replication number q, therefore, the sum of squares due to the main effect A is: $SSA = \frac{[A]^2}{4q}$ with degree of freedom 1

Similarly, $SSB = \frac{[B]^2}{4q}$ with degree of freedom 1

And, $SSAB = \frac{[AB]^2}{4q}$ with degree of freedom 1

Mean sum of squares due to factor A, B and AB can be obtained by dividing the corresponding sum of squares by the degree of freedom

$MSA = SSA/1 = \frac{[A]^2}{4q}$

$MSB = SSB/1 = \frac{[B]^2}{4q}$

$MSAB = SSAB/1 = \frac{[AB]^2}{4q}$

Hence, the test for significance of any factorial may be obtained by using the test statistic

$$F = \frac{Mean\ Sum\ of\ Square\ due\ to\ effect\ A}{MSE}$$

Where MSE be the error mean sum of squares for the analysis of corresponding design and

$$F \sim F_{[(1,3q-1)],\alpha}$$

The hypothesis of the presence of factorial effect is rejected at level of α if $F > F_{[(1,3q-1)],\alpha}$, otherwise null hypothesis is accepted.

ANOVA table for 2^2 experiments with q randomized blocks

Source of Variation	***Degree of Freedom***	***Sum of Squares***	***Mean Sum of Square***	***Variance Ratio***	
				$F_{Cal.}$	***$F_{Tab.}$***
Blocks	q – 1	SSC (Blocks)	MS (Blocks)		
Main effect A	1	SSA	MSA	$F_A = \frac{MSA}{MSE}$	$F_{\{1,3(q-1), \alpha\}}$
Main effect B	1	SSB	MSB	$F_B = \frac{MSB}{MSE}$	$F_{\{1,3(q-1), \alpha\}}$
Interaction (AB)	1	SS(AB)	MS(AB)	$F_{AB} = \frac{MS(AB)}{MSE}$	$F_{\{1,3(q-1), \alpha\}}$
Error	3(q-1)	SSE	MSE		
Total	4q-1	TSS			

Yates Method of computing factorial effects total

Yates give a systematic method of obtaining the various effect total for any 2^m experiments without writing down the algebraic expression. The steps are:

1. First write down the four treatments combinations systematically in the first column, starting with the treatment combination (1) and then introducing the letter A, B in terms after introducing a letter write down its combination with all the previous treatment combinations and then introduce a new letter. Repeat this until all the letters have been exhausted.

2. Next, write down the treatment total from all the replicate, in the second column against the appropriate treatment combinations.
3. The first two columns we get from the observed data. for obtaining column three we break the even number of values in the second columns into consecutive pairs. Then in the first half of the third column, we write down the sums of the values in these pair in order and in 2^{nd} half of the third column we write down in order the difference of the values in the pairs in the second column (the first member subtracted from the second member of the pair).
4. We next break the values in the third column into consecutive pairs and put the sum and difference of the members of these pair in order of the fourth column.

Yates Method of Computing Factorial Effect for a 2^2 experiment

Treatment (1)	***Total (2)***	***(3)***	***(4)***	***Effect (5)***
1	[1]	[1] + [a]	[1] + [a] + [b] + [ab]	[1]
a	[a]	[b] + [ab]	[a] - [1] + [ab] - [b]	[A]
b	[b]	[a] - [1]	[b] + [ab] - [1] - [a]	[B]
ab	[ab]	[ab] - [b]	[ab] - [b] - [a] + [1]	[AB]

Example: An experiment was planned to study the effect of sulphate of potash and superphosphate on the yield of potatoes. All the combinations of 2 levels of superphosphate [0 cent (p_0) and 5 cent (p_1)/acre] and two levels of sulphate of potash [0 cent (k_0) and 5 cent (k_1)/acre] were studied in a randomized block design with 4 replications for each. The (1/70) yields [lb. per plot = (1/70) acre] obtained are given below:

Block	***Yields (lbs per plot)***			
I	(1)	k	p	kp
	23	25	22	38
II	p	(1)	k	kp
	40	26	36	38
III	(1)	k	pk	p
	29	20	30	20
IV	kp	k	p	(1)
	34	31	24	28

Analyze the data and give your conclusion.

Solution: Taking deviation from y =29, we re-arrange the data in the above table for computations of S.S. due to treatments and blocks:

Treatment Combination	***Blocks***				***Treatment Totals (T_i)***	T_i^2
	I	***II***	***III***	***IV***		
(1)	-6	-3	0	-1	-10	100
k	-4	7	-9	2	-4	16
p	-7	11	-9	-5	-10	100
kp	9	9	1	5	24	576
Block Totals (B_j)	-8	24	-17	1	G = 0	
B_j^2	64	576	289	1		

H_0: The data is homogeneous with respect to the blocks and the treatments.

N = $4 \times 4 = 16; G = 0; R.S.S. = \sum_i \sum_j y_{ij}^2 = 660$

C.F. = Correction Factor = $\frac{G^2}{N} = \frac{(0)^2}{16} = 0$

Total S.S. = RSS-C.F. = 660-0 = 660

Block S.S.= $\frac{1}{4}\sum_i B_j^2 - C.F. = \frac{64+576+289+1}{4} = \frac{930}{4} = 232.50$

Treatment S.S. = $\frac{1}{4}\sum_i T_i^2 - C.F. = \frac{100+16+100+579}{4} = \frac{792}{4} = 198$

Error S.S.= 660 - (232.50+198.0) = 229.50

We now compute the factorial effect totals by Yates Method.

Yates' Method for 2^2 Experiment

Treatment Combination (1)	***Total Yield from all blocks (2)***	***(3)***	***Factorial effects totals (4)***	***S.S. (5) = $\frac{(4)^2}{4r}$***
'1'	-10	-14	0 = G	$(0)^2/16$=0=C.F.
k	-4	14	40 = [K]	$(40)^2/16$ =100=S_K^2
p	-10	6	28 = [P]	$(28)^2/16$ = 49 = S_P^2
kp	24	34	28 = [KP]	$(28)^2$ /16 = 49 =S_{KP}^2

ANOVA Table for 2^2 Experiment

Source of Variation	***Degree of Freedom***	***Sum of Squares***	***Mean Sum of Square***	***Variance Ratio***	
				$F_{Cal.}$	$F_{Tab.}$
Blocks	3	232.5	77.5	3.04	3.86
Treatments	3	198.0	66.0	2.59	3.86

K	1	100	100	3.92	5.12
P	1	49	49	1.92	5.12
KP	1	49	49	1.92	5.12
Error	9	229.5	25.5	-	-
Totals	15	660	-	-	-

As in each of the cases, the computed value of F is less than the corresponding tabulated (critical) value, there are no significant main or interaction effects present in the experiment. The blocks as well as treatments do not differ significantly.

Since the blocks do not differ significantly, we conclude that there is no special contribution from fluctuations in soil fertility. Thus, the division of the whole experimental area into blocks does not result in any gain in accuracy.

Remark. It may be noted that $S_K^2 + S_P^2 + S_{KP}^2 = 100 + 49 + 49 = 198 \ = \text{Treatment S.S.}$, as it should be.

5.4.2 2^3 Factorial Experiment

Here we have three factors, A, B and C, each have two levels. So, the treatment combinations in this case will be 8 and can be written in the systematic order as: 1, a, b, ab, c, ac, bc, abc. These 8 treatment combinations can be compared by using any of design as CRD, RBD and LSD. In this experiment, there are three main effects, A, B and C and three first order interaction effects AB, AC and BC and one second order interaction effect ABC.

Main effects and interaction effects

Let us consider factor A, then the effect of factor A has four simple effects:

i. The effect of changing factor A from its first level to second in presence of first level of factor B and C is (a) – (1).
ii. The effect of changing factor A from its first level to second in presence of second level of factor B and first level of C is (ab) – (b).
iii. The effect of changing factor A from its first level to second in presence of first level of factor B and second level of C is (ac) – (c).
iv. The effect of changing factor A from its first level to second level in the presence of second level of B and C is (abc) – (bc).

Similarly, the effects of factor B and C is:

	For B	***For C***

(i)	(b) – (1)	(c) – (1)
(ii)	(bc) – (c)	(ac) – (a)
(iii)	(ab) – (a)	(bc) – (b)
(iv)	(abc) – (ac)	(abc) – (ab)

Now, the main effect of factor A is given by the average of simple effects

$$A = \frac{1}{4}[(a) - (1) + (ab) - (b) + (ac) - (c) + (abc) - (ab)]$$

$$A = \frac{1}{4}(a - 1)(b + 1)(c + 1)$$

Similarly for B and C:

$$B = \frac{1}{4}[(abc) + (bc) - (ac) - (c) + (ab) + (b) - (a) - (1)]$$

$$B = \frac{1}{4}(a + 1)(b - 1)(c + 1)$$

$$C = \frac{1}{4}[(abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)]$$

$$C = \frac{1}{4}(a + 1)(b + 1)(c - 1)$$

The interaction of A with B can be obtained as:

AB (when C is first level) $= \frac{1}{2}[(ab) - (b) - (a) + (1)]$

AB (when C is second level) $= \frac{1}{2}[(abc) - (bc) - (ac) + (c)]$

Finally, the average of these two gives the interaction effect AB.

$$AB = \frac{1}{4}[(abc) - (bc) - (ac) + (c) + (ab) - (b) - (a) + (1)]$$

$$AB = \frac{1}{4}(a - 1)(b - 1)(c + 1)$$

Similarly, we can obtain the other two effects AC and BC as:

$$AC = \frac{1}{4}[(abc) - (bc) + (ac) - (c) - (ab) + (b) - (a) + (1)]$$

$$AC = \frac{1}{4}(a - 1)(b + 1)(c - 1)$$

And

$$BC = \frac{1}{4}[(abc) + (bc) - (ac) - (c) - (ab) - (b) + (a) + (1)]$$

$\text{BC} = \frac{1}{4}(a+1)(b-1)(c-1)$

The second order interaction of three factors A, B, and C is given by the half of the difference AB (C at first level) and AB (C at second level). Thus:

$\text{ABC} = \frac{1}{4}[(abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)]$

$\text{ABC} = \frac{1}{4}(a-1)(b-1)(c-1)$

Table for signs and divisors given M, A, B, AB, C, AC, BC and ABC in terms of treatments means is given as follows:

Effects	***Treatment Means***								***Divisors***
	(1)	***(a)***	***(b)***	***(ab)***	***(c)***	***(ac)***	***(bc)***	***(abc)***	
M	+	+	+	+	+	+	+	+	8
A	-	+	-	+	-	+	-	+	4
B	-	-	+	+	-	-	+	+	4
AB	+	-	-	+	+	-	-	+	4
C	-	-	-	-	+	+	+	+	4
AC	+	-	+	-	-	+	-	+	4
BC	+	+	-	-	-	-	+	+	4
ABC	-	+	+	-	+	-	-	+	4

Analysis of Variance of 2^3 Factorial Experiment

To obtain sum of square due to effects, we need treatment totals:

[A] = [ABC] – [BC] + [AC] – [C] + [AB] – [B] + [A] – [1]

[B] = [ABC] + [BC] - [AC] – [C] + [AB] + [B] - [A] – [1]

[C] = [ABC] - [BC] - [AC] + [C] + [AB] - [B] - [A] + [1]

[AC] = [ABC] - [BC] + [AC] - [C] - [AB] + [B] - [A] + [1]

[BC] = [ABC] + [BC] - [AC] + [C] - [AB] - [B] + [A] + [1]

[ABC] = [ABC] + [BC] - [AC] + [C] - [AB] + [B] + [A] - [1]

Now, the sum of square due to factors A, B, AB, C, AC, BC and ABC are given by:

$\text{SSA} = \frac{[A]^2}{8q}$ $\quad$ $\text{SSB} = \frac{[B]^2}{8q}$ $\quad$ $\text{SSC} = \frac{[C]^2}{8q}$

$\text{SS}(AC) = \frac{[AC]^2}{8q}$ $\quad$ $\text{SS}(BC) = \frac{[BC]^2}{8q}$ $\quad$ $\text{SS}(ABC) = \frac{[ABC]^2}{8q}$

Where, q is the blocks size as replications numbers of each treatment.

Now, the test for significance of any factorial effect i.e., main effect and interaction effect can be obtained by using the quantity.

$$F_A \frac{Mean\ Sum\ of\ Square\ due\ to\ effect\ A}{MSE} \sim F_{(1,7q-1),\alpha}$$

And similarly, we can obtain test of significance of other factors.

Hence, the null hypothesis of absence of factorial effect is reject at the level of significance α*100% if $F_A > F_{(1,7q-1),}$ otherwise null hypothesis is accepted.

ANOVA Table for 2^3 Experiment with q randomized blocks

Source of Variation	***Degree of Freedom***	***Sum of Squares***	***Mean Sum of Square***	***Variance Ratio***	
				$F_{Cal.}$	$F_{Tab.}$
Blocks	q – 1	SSC (Blocks)	MS (Blocks)	F_{Blocks}	$F_{\{(q-1),7(q-1),\alpha\}}$
A	1	SSA	MSA	F_A	$F_{\{1,7(q-1),\alpha\}}$
B	1	SSB	MSB	F_B	$F_{\{1,7(q-1),\alpha\}}$
AB	1	SS(AB)	MS(AB)	F_{AB}	$F_{\{1,7(q-1),\alpha\}}$
C	1	SSC	MSC	F_C	$F_{\{1,7(q-1),\alpha\}}$
AC	1	SS(AC)	MS(AC)	F_{AC}	$F_{\{1,7(q-1),\alpha\}}$
BC	1	SS(BC)	MS(BC)	F_{BC}	$F_{\{1,7(q-1),\alpha\}}$
ABC	1	SS(ABC)	MS(ABC)	F_{ABC}	$F_{\{1,7(q-1),\alpha\}}$
Error	7(q-1)	SSE	MSE		
Total	8q-1	TSS			

Yate's Method for computing Factorial Effect Total for a 2^3 experiment

We follow the instructions given in case of 2^3 experiment and obtain one more column as shown below:

Treatment	***Yield***				***(6)***
(1)	***(2)***	***(3)***	***(4)***	***(5)***	
1	(1)	$u_1 = (1) + (a)$	$v_1 = u_1 + u_2$	$w_1 = v_1 + v_2$	[1]
a	(a)	$u_2 = (b) + (ab)$	$v_2 = u_3 + u_4$	$w_2 = v_3 + v_4$	[A]
b	(b)	$u_3 = (c) + (ac)$	$v_3 = u_5 + u_6$	$w_3 = v_5 + v_6$	[B]
ab	(ab)	$u_4 = (bc) + (abc)$	$v_4 = u_7 + u_8$	$w_4 = v_7 + v_8$	[AB]

c	(c)	$u_5 = (a) + (1)$	$v_5 = u_2 + u_7$	$w_5 = v_2 + v_1$	[C]
ac	(ac)	$u_6 = (ab) + (b)$	$v_6 = u_4 + u_3$	$w_6 = v_4 + v_3$	[AC]
bc	(bc)	$u_7 = (ac) + (c)$	$v_7 = v_6 + u_5$	$w_7 = v_6 + v_5$	[BC]
abc	(abc)	$u_8 = (abc) + (bc)$	$v_8 = u_8 + u_7$	$w_8 = v_8 + v_7$	[ABC]

Example: A 2^3 factorial design was used to develop a nitride etch process on a single-wafer plasma etching tool. The design factors are the gap between the electrodes, the gas flow (C_2F_6 is used as the reactant gas), and the RF power applied to the cathode. Each factor is run at two levels, and the design is replicated twice. The response variable is the etch rate for silicon nitride (Å/m). The etch rate data are shown in table below:

Replicate-1

'1' 550	(abc) 729	(b) 633	(ac) 749
(c) 1037	(ab) 642	(bc) 1075	(a) 669

Replicate-2

'1' 604	(c) 1052	(bc) 1063	(ab) 635
(a) 650	(ac) 868	(b) 601	(abc) 860

Solution:

Coded Factors			***Etch Rate***		***Total***
A	***B***	***C***	***Replicate-1***	***Replicate-2***	
0	0	0	550	604	(1) = 1154
1	0	0	669	650	a = 1319
0	1	0	633	601	b = 1234
1	1	0	642	635	ab = 1277
0	0	1	1037	1052	c = 2089
1	0	1	749	868	ac = 1617
0	1	1	1075	1063	bc = 2138
1	1	1	729	860	abc = 1589

Using the totals under the treatment combinations as shown above, the estimate of factor effects can be obtained as follows:

$$A = \frac{1}{4n}[a - (1) + ab - b + ac - c + abc - bc]$$

$$= \frac{1}{8}[1319 - 1154 + 1277 - 1234 + 1617 - 2089 + 1589 - 2138]$$

$$A = -101.625$$

$B = \frac{1}{4n}[b + ab + bc + abc - (1) - a - c - ac]$

$= \frac{1}{8}[1234 + 1277 + 2138 + 1589 - 1154 - 1319 - 2089 - 1617]$

$B = 7.375$

$C = \frac{1}{4n}[c + ac + bc + abc - (1) - a - b - ab]$

$= \frac{1}{8}[2089 + 1617 + 2138 + 1589 - 1154 - 1319 - 1234 - 1277]$

$B = 306.125$

$AB = \frac{1}{4n}[ab - a - b + (1) + abc - bc - ac + c]$

$= \frac{1}{8}[1277 - 1319 - 1234 + 1154 + 1589 - 2138 - 1617 + 2089]$

$AB = -24.875$

$AC = \frac{1}{4n}[(1) - a + b - ab - c + ac - bc + abc]$

$= \frac{1}{8}[1154 - 1319 + 1234 - 1277 - 2089 + 1617 - 2138 + 1589]$

$AC = -153.625$

$BC = \frac{1}{4n}[(1) + a - b - ab - c - ac + bc + abc]$

$= \frac{1}{8}[1154 + 1319 - 1234 - 1277 - 2089 - 1617 + 2138 + 1589]$

$BC = -2.125$

$ABC = \frac{1}{4n}[abc - bc - ac + c - ab + b + a - (1)]$

$= \frac{1}{8}[1589 - 2138 - 1617 + 2089 - 1277 + 1234 + 1319 - 1154]$

$ABC = 5.625$

The largest effects are for power (C = 306.125), gap (A = -101.625) and the power-gap interaction (AC = -153.625).

The sums of squares are calculated as follows: $SS = \frac{Constrast^2}{8n}$

$SS_A = \frac{(-813)^2}{16} = 41{,}310.5625$ $SS_{AB} = \frac{(-199)^2}{16} = 2475.0625$

$SS_B = \frac{(59)^2}{16} = 217.5625$ $SS_{AC} = \frac{(-1229)^2}{16} = 94{,}402.5625$

$SS_C = \frac{(2449)^2}{16} = 374{,}850.0625$ $SS_{BC} = \frac{(-17)^2}{16} = 18.0625$

$SS_{ABC} = \frac{(45)^2}{16} = 126.5625$

The total sum of square is TSS = 531,420.9375 and by subtraction, SSE = 18,020.50. The ANOVA table can now be used to confirm the magnitude of these effect.

ANOVA Table

Source of Variation	***Degree of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$F_{Cal.}$	$F_{Tab.}$
Gap (A)	1	41,310.5625	41,310.5625	18.34	5.3177
Gas Flow (B)	1	217.5625	217.5625	0.10	5.3177
Power (C)	1	374,850.0625	374,850.0625	166.41	5.3177
AB	1	2475.0625	2475.0625	1.10	5.3177
AC	1	94,402.5625	94,402.5625	41.91	5.3177
BC	1	18.0625	18.0625	0.01	5.3177
ABC	1	126.5625	126.5625	0.06	5.3177
Error	8	18,020.5000	2252.5625		
Total	15	531,420.9375			

From the above ANOVA table, it can be noted that the main effects of Gap and Power are highly significant (both have very large $F_{Tab.}$ values). The AC interaction is also highly significant, thus, there is a strong interaction between Gap and Power.

5.5 3^n Factorial Experiments

When factors are taken at 3 levels instead of 2, the experiment becomes more informative. Let the n-factor be denoted by A, B, C, ……. and so on, each having three levels. The levels of these factors are denoted by 0, 1 and 2. The possible combinations of these three levels of each factor will give rise to 3^n treatment combination, each being n-tuple like (x_1, x_2, ……., x_n), where x_1 be the level of first factor, x_2 be the level of second factor and so on.

Here, we use the number system reduce modulo 3; i.e.:

0 = 3 = 6 = 9 = …………….

1 = 4 = 7 = 10 = …………….

2 = 5 = 8 = 11 = …………….

In this system, we divide a number greater than or equal to 3 by 3 and take the remainder to be equal to the original number.

Standard order of Treatment Combinations

The complete list of 3^n treatment combinations can be conveniently written in standard order as follows:

Factorial Design	Factors	Order
3^1	A	(1), a_1, a_2
3^2	A and B	(1) a^1, a^2, b^1, a^1b_1, a_2b_1, b_2, a_1b_2, a_2b_2
3^3	A, B and C	(1), a_1, a_2, b_1, a_1b_1, a_2b_1, a_2b1, b_2, a_1b_2, a_2b_2, c_1, a_2c_1, b_1c_1, $a_1b_1c_1$, $a_2b_1c_1$, b_2c_1, $a_1b_2c_1$, $a_2b_2c_1$, c_2, a_1c_2, a_2c_2, b_1c_2, $a_1b_1c_2$, a_1, b_1c_2, b_2c_2, $a_1b_2c_2$, $a_2b_2c_2$

5.5.1 3^2 Factorial Experiment

In this experiment there are two factors say A and B each having 3 levels 0 1 and 2. Therefore, these are 9 treatment combinations of type (x_1, x_2) where x_1 and x_2 can take only of the values 0 1 and 2. Thus the 9-treatment combination are

0 0, 0 1, 0 2, 1 0, 1 1, 1 2, 2 0, 2 1, 2 2

Among these 9 treatment combinations there will be 8 comparisons (degree of freedom =8) can be partitioning as 2 degrees of freedom for each of main effects A and B and 4 degrees of freedom for interaction A $\times B$. This combination (component) can be denoted by a 2-way table as follows:

		Levels of A			
Levels of B		*0*	*1*	*2*	$[\text{B}]_0$
	0	0 0	0 1	0 2	$[\text{B}]_1$
	1	1 0	1 1	1 2	$[\text{B}]_2$
	2	2 0	2 1	2 2	
		$[A]_0$	$[A]_1$	$[A]_2$	

Analysis

The sum of squares can be obtained as

$$\text{SSA} = \frac{\left([A]_0^2+[A]_1^2+[A]_2^2\right)}{3q} - \frac{G^2}{9q}$$

$$\text{SSB} = \frac{([B]_0^2+[B]_1^2+[b]_2^2)}{3q} - \frac{G^2}{9q}$$

$$\text{SS(AB)} = \frac{(00)^2+(01)^2+\ldots\ldots\ldots(21)^2+(22)^2}{q} - \frac{G^2}{9q} - SSA - SSB$$

Where q is block.

Four degrees of freedom of AB can be further partition into 2 more orthogonal components are very useful in confounding in 3^2 factorial experiment. Thus, the interaction $A \times B$ carrying 4 degrees of freedom can be partitioned AB and AB^2as

AxB	AB gives $x_1+x_2 = 0, 1, 2$	Modulo 3
	AB^2 gives $x_1+2x_2 = 0, 1, 2$	Modulo 3

Defining the equation, dividing the 9-treatment total intro 3 groups, A comparison between which gives the corresponding components. Thus

$x_1 + x_2 = 0$ gives $[00] + [12] + [21] = [\text{AB}]_0$

$x_1 + x_2 = 1$ gives $[01] + [10] + [22] = [\text{AB}]_1$

$x_1 + x_2 = 2$ gives $[02] + [20] + [11] = [\text{AB}]_2$

Similarly,

$x_1 + 2x_2 = 0$ gives $[00] + [12] + [21] = [\text{AB}^2]_0$

$x_1 + 2x_2 = 1$ gives $[02] + [21] + [10] = [\text{AB}^2]_1$

$x_1 + 2x_2 = 2$ gives $[01] + [12] + [20] = [\text{AB}^2]_2$

And hence sum of square due to component AB is

$$\text{SS(AB)} = \frac{([AB]_0^2+[AB]_1^2+[AB]_2^2)}{3q} - \frac{G^2}{9q}$$

And for component AB^2 is

$$\text{SS(AB}^2) = \frac{([AB^2]_0^2+[AB^2]_1^2+[AB^2]_2^2)}{3q} - \frac{G^2}{9q}$$

The first two columns of ANOVA table are given as:

Source of Variation	***Degree of Freedom***

Replication		q-1
A		2
B		2
AB	AB	2
	AB^2	2
Error		8(q-1)
Total		9q-1

5.5.2 3^3 Factorial Experiment

In this experiment, we have 27 treatment combinations of the form (x_1, x_2 x_3) where x_1 , x_2 and x_3 are the levels of A, B and C respectively. All the treatment combination can systematically be written as:

A	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
0	0	0	0	0	1	0	0	2
1	0	0	1	0	1	1	0	2
0	1	0	0	1	1	0	1	2
1	1	0	1	1	1	1	1	2
2	0	0	2	0	1	2	0	2
2	1	0	2	1	1	2	1	2
0	2	0	0	2	1	0	2	2
1	2	0	1	2	1	1	2	2
2	2	0	2	2	1	2	2	2

The 27-treatment combination will have a sum of all squares carrying 26 degrees of freedom. The treatment sum of squares can be calculated from 27 treatment totals taking over the q replicates. In these experiments the treatments can be sub divided into the main effect and intersection effects having degrees of freedom as follows:

Source of Variation	***Degree of Freedom***
Replication	q-1
Treatment	26
A	2
B	2

C	2
A×B	4
A×C	4
B×C	4
A×B×C	8
Error	26(q-1)
Total	27q-1

The sum of squares due to main effects and two factor interactions are calculated from the three-two way in the usual manners. However, the three-factor interaction (A×B×C) sum of square is obtained by the subtractions of these components from treatment sum of square. Here each set of 4 or 8 degrees of freedom can also be partitioned into their orthogonal component carrying 2 degrees of freedom as follows:

AxB	AB gives $x_1+x_2 = 0, 1, 2$	Modulo 3
	AB^2 gives $x_1+2x_2 = 0, 1, 2$	Modulo 3
AxC	AC gives $x_1+x_3 = 0, 1, 2$	Modulo 3
	AC^2 gives $x_1+2x_3 = 0, 1, 2$	Modulo 3
AxC	BC gives $x_2+x_3 = 0, 1, 2$	Modulo 3
	BC^2 gives $x_2+2x_3 = 0, 1, 2$	Modulo 3
AxBxC	ABC gives x1+x2+x3 = 0, 1, 2	Modulo 3
	AB^2C gives x1+2x2+x3 = 0, 1, 2	Modulo 3
	ABC^2 gives x1+x2+2x3 = 0, 1, 2	Modulo 3
	AB^2C^2 gives x1+x2+x3 = 0, 1, 2	Modulo 3

Using each of the defining equations we divide the 27 treatment combinations intro 3 groups and comparison among these groups total is the corresponding sum of square carrying 2 degrees of freedom.

Example: The sum of square due to the component AB^2C can be obtained as follows:

$$SS(AB^2C) = \frac{[AB^2C]_0+[AB^2C]_1+[AB^2C]_2}{3q} - \frac{G^2}{27q}$$

Where,

$$\left[AB^2C\right]_0 = [000] + [011] + [110] + [212] + [201] + [102] + [121] + [022] + [220]$$

$$\left[AB^2C\right]_1 = [001] + [100] + [012] + [111] + [202] + [210] + [020] + [122] + [221]$$

$\left[AB^2C\right]_2 = [002] + [101] + [010] + [112] + [200] + [211] + [021] + [120] + [222]$

Similarly, the other sum of square can be obtained by using their defining equations.

Each mean sum of squares can be obtained by dividing the corresponding sum of squares by its degree of freedom. The mean sum of square due to error can be obtained in the usual way.

ANOVA Table for a 3^3 Factorial Experiment

The first two column of the ANOVA table as follows:

Source of Variation		***Degree of Freedom***
Replications		q-1
Treatment		26
A		2
B		2
C		2
$A \times B$	AB	2
	AB^2	2
$A \times C$	AC	2
	AC^2	2
$B \times C$	BC	2
	BC^2	2
$A \times B \times C$	ABC	2
	AB^2C	2
	ABC^2	2
	AB^2C^2	2
Error		26(q-1)
Total		27q-1

5.6 Self-Assessment Exercise

Question-1: Analyze the following factorial design

Replicate 1							
nk	kd	1	nd	d	k	n	nkd
291	398	101	373	312	245	106	450

Replicate 2							
kd	d	n	nk	k	nkd	nd	1
907	329	89	306	272	449	338	106
Replicate 3							
d	1	nk	nkd	nd	k	n	kd
323	87	334	471	324	279	128	423
Replicate 4							
nd	nk	k	d	n	1	nkd	kd
361	272	302	324	103	131	437	445

Question-2: The following table gives the number of seeds germinated (out of 10 seeds) in each of 16 plots along with the treatments applied. Here, rows correspond to different levels of temperature and columns to different depths of sowing:

Temperature	*Depth of sowing*			
1	np (7)	n (3)	1 (2)	p (8)
2	n (4)	p (4)	np (6)	1 (3)
3	1 (2)	np (8)	p (3)	n (4)
4	p (5)	1 (3)	n (4)	np (8)

5.7 Summary

This unit gives a complete overview of the factorial experiments, covering both 2^n and 3^n factorial experiments, their various types such as 2^2, 2^3, 3^2 and 3^3 along with their analysis.

5.8 References

- Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer-Verlag, New York.
- Gupta S.C., Kapoor V.K. (2007). *Fundamentals of Applied Statistics*, Sultan Chand & Sons.
- Gupta S.P. (2021). *Statistical Methods*, Sultan Chand & Sons.

- Gupta V.K. & Nigam A.K. Nigam (1978-79). *Handbook an analysis of Agriculture Experiment*, IASRI Publication.

5.9 Further Reading

- Cochran, W. G. and Cox, G. M. (1957). *Experimental Design*. John Wiley & Sons, New York.
- Cochran, W.G. and Cox, G.M. (2003). *Experimental Design*, Asia Publishing House
- Das, M. N. and Giri, N. S. (1986). *Design and Analysis of Experiments (2nd Edition).* Wiley.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*. New Age International (P) Ltd. New Delhi.
- Montgomery, D.C. (2017). *Design and Analysis of Experiments (9th Edition)*. John Wiley & Sons.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments,* Wiley, New York.

UNIT:6 CONFOUNDING

Structure

6.1 Introduction

The 2^2 and 2^3 factorial experiments should be conducted by using completely randomized, randomized block or Latin square designs. These experiments can also be analysed by breaking the treatment components into main effect and interaction components. The next factorial, 2^4 has 16 treatment combinations and it is not advisable to adopt a randomized block design for it, because blocks of 16 plots are too big to ensure homogeneity within them. A new device is, therefore, necessary for designing experiments with a large number of treatments. One such device is to take blocks of size less than the number of treatments and have more than one block per replication. The treatments are then divided into as many groups as the number of blocks per replication. The different groups of treatments are allotted to the blocks.

Example: We can take for 24 factorial two blocks each of eight units and divide the 16 treatment combinations into two groups of each for allotment to the two blocks.

In general, the block size for 2n factorials is of form 2r. There are many ways of grouping the treatments into as many groups as the number of blocks per replication. For obtaining interaction contrast the treatment combinations are divided into two groups. Two such groups representing a suitable interaction, say P, can be taken to form the contents of two blocks, each containing half the total number of treatments.

In such cases the contrast of the interaction P and the contrast between the two blocks totals are given by the same function. They are, therefore, mixed up with block effects and cannot be separated. In other words, the interaction P, has been confounded with the blocks. Evidently, P has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by making one or more interaction contrasts identical with block contrasts, is known as Confounding. Though we have introduced the concept of confounding by having only two blocks, the number of blocks can be any power of 2.

Preferably, only higher order interaction, that is, interactions with three or more factors are confounded, because their loss is immaterial. As an experimenter is more interested in main effects and two factor interactions, these should not be confounded as far as possible.

The designs for such confounded factorials can be called incomplete randomized block design. The treatment groups are first allotted at random to the different blocks. The treatments allotted to a block are then distributed at random to its different units.

6.2 Objectives

After going through this unit, you should be able to:

- Know about the concept of orthogonality in the design of experiment
- Understand the concept of confounding in factorial experiments
- Understand the concept of partially confounded in factorial experiments

6.3 Orthogonality

Suppose that we have a random sample of n independent observation $y_1, y_2 \dots\dots y_n$ from a normal population with variance σ^2 .let us consider two orthogonal contrast $A = \sum_{i=1}^{n} \mu \times y_i$ and $B = \sum_{i=1}^{n} \mu_i \times y_i$ with $\sum_{i=1}^{n} \lambda_i = 0$ and $\sum_{i=1}^{n} \mu_i = 0$.

Therefore, A and B are independent and we can use A & B to estimate two different effects because the error in two estimates will not be related. These estimates are also said to be orthogonal. The orthogonality of a design ensures that the different effects will be capable of separate estimation in testing without any difficulties, hence if data arise from an orthogonal design, then there will be no difficulty of independent estimation and test effects.

The designs CRD, RBD, and LSD give us orthogonal design. The $2^2$and 2^3 etc. Factorial experiments are conducted by using CRD, RBD and LSD. However, the difficulty in conducting the factorial experiment in this design is that as the number of factors or the number of levels of the factor or both increases, the number of treatment combinations to be compared increases rapidly. This results in the use of large-size blocks or squares to accommodate all the treatment combinations.

Example: In an 2^4 factorial experiment, there are 16 treatment combinations and it is advisable to adopt RBD for it because blocks of 16 plots are too big, here to ensure homogeneity within them. Therefore, it is necessary to have a new device for designing experiments with a large number of treatments. Once such a device takes blocks of size less than the number of treatments and has more than one block, then the treatments are divided into blocks per replication.

The different groups of treatments are allotted to the blocks in such a way that the only unimportant treatment combination gets mixed up with the block comparison. These treatment comparisons are said to be confounded or mixed up with block effects.

These effects cannot be estimated for testing separately. However, the remaining treatment effects, which are not confounding by the block effect, are still capable of separate estimation in testing. In a confounding design, we lose information on some treatment comparisons (completely or partially) which are confounding. Therefore, there should be least important comparison and generally we choose highest order interaction in confounding.

6.4 Confounding in Factorial Experiment

Confounding in experiment designs is an arrangement of treatment combinations in blocks in which less important treatment effects are purposely confounded with the blocks.

There are two types of confounding: -

1. Complete confounding
2. Partial confounding

When there are two or more replications, then the question arises whether we confounded the same interaction in each replication or different sets of interactions in different replications. If the same set of interactions is confounded in all the replications, then confounding is called complete confounding. In complete confounding, all the information on confounded interaction is lost and we lose all the information from all the replications.

If different interaction steps are confounded in different replications, then confounding is called partial confounding. In this confounding, the confounded interaction can be recovered from those replications in which they are not confounded.

6.4.1 Confounding in 2^3 Factorial Experiment

In this experiment, we have eight treatment combinations under comparison and suppose we decide to use two blocks to accommodate a treatment combination is replicated. Thus, we have four plots of each. We will divide the 8 treatment combinations each and allocate the two groups to the two blocks at random. Here it is to be notable that 2^n factorial experiment conducted $2^{(k-1)}$ treatment combinations in which $2^{(k)} - 1 - k$ treatment combination will be generalized. They are automatically confounded with the block effects. if k=1, then only one treatment combination will be confounded and if k>1, then more than one treatment combination will be confounded according to the above rule. Therefore, in this case k=1, only one treatment combination will be confounded and we decide highest order interaction ABC to be confounded.

The interaction ABC is depended on $(abc) + (a) + (b) + (c) - (ab) - (ac) - (bc) - 1.$

Let us apply the four treatment combinations with the (+) sign in one block and the remaining 4 with the (-) sign in the second block. Then, block 1^{st} and 2^{nd} contain the treatment combination as:

Block I	***Block II***
(abc)	(ab)
(a)	(ac)
(b)	(bc)
(c)	(1)

Here, the contrast measuring the interaction ABC also contain block effect, i.e., the effect of block 1^{st} - effect of block $2^{nd,}$ and we say that the interaction ABC is mixed up or confounded with block effect and we loss information ABC. However, the other six contrasts of treatment combinations A, B, AB, C, AC, and BC still maintain their orthogonality to the block as each treatment combination from block 1st (block 2nd) with (+) sign and two treatment combinations with (-) sign of the remaining six treatment combinations therefore they will contain no block effect.

Thus, in this allocation to two blocks, there is no difficulty in the estimation or testing of the remaining six treatment combinations.

The first two columns of the ANOVA table for 2^3 factorial experiment is given by:

Source of Variation	***Degree of Freedom***
Blocks	2q-1
A	1
B	1
AB	1
C	1
AC	1
BC	1
Error	6(q-1)
Total	8q-1

Example: Analyze the following 2^3 completely confounded factorial design:

Replicate-1

Block I	'1' 101	(nk) 291	(np) 373	(kp) 391
Block II	(nkp) 450	(n) 106	(k) 265	(p) 312

Replicate-2

Block III	'1' 106	(nk) 306	(np) 338	(kp) 407
Block IV	(nkp) 449	(n) 89	(k) 272	(p) 324

Replicate-3

Block V	'1' 87	(nk) 334	(np) 324	(kp) 423
Block VI	(nkp) 471	(n) 128	(k) 279	(p) 323

Replicate-4

Block VII	'1' 131	(nk) 272	(np) 361	(kp) 445
Block VIII	(nkp) 437	(n) 103	(k) 302	(p) 324

(N = Nitrogen; P= Phosphate; K = Potash)

Solution: Since in the above 2^3 factorial experiment the replicate has been divided into blocks of 4 plots each, it is a 2^3 confounded design. A careful examination of the treatment combinations in different blocks reveals that the interaction NPK has been confounded in each replicate. [Note that in each replicate, the treatment combinations in the block containing T have no or an even number' of treatments common with npk.]. Hence the above design is a 2^3 factorial with the interaction NPK completely confounded with blocks.

The S.S. due to the six unconfounded factorial effects, viz., the main effects N, P and K and the first order interactions NP, KP and NK are obtained by Yates' technique as usual.

Yates' Method for Factorial Effects and S.S.

Treatment	***Totals***	***1***	***2***	***3***	***Effects***	***S.S.= [Effect Totals)²/32***
I	425	851	3,172	9,324	G	C.F. = 27,16,780.5
n	426	2,321	6,152	340	[N]	$S_N^2 = 3{,}612.5$
k	1,118	2,679	86	2,264	[K]	$S_k^2 = 1{,}60{,}178.0$
nk	1,203	3,473	254	112	[NK]	$S_{NK}^2 = 392.0$
p	1,283	1	1,470	2,980	[P]	$S_P^2 = 2{,}77{,}512.5$
np	1,396	85	794	168	[NP]	$S_{NP}^2 = 882.0$
kp	1,666	113	84	-676	[KP]	$S_{KP}^2 = 14{,}280.5$
nkp	1,807	141	28	-56	[NKP]	Not estimable

∴ S.S. due to treatments = $S_N^2 + S_K^2 + S_P^2 + S_{NP}^2 + S_{NK}^2 + S_{KP}^2 = 456857.5$

(Since NPK is completely confounded with blocks, its effects enter into the error S.S.)

$R.S.S. = 31{,}82{,}118.0$

$C.F. = \frac{G^2}{8\times4} = \frac{(9{,}324)^2}{32} = 27{,}16{,}780.5$

Total S.S. = R.S.S.-C.F. = 31,82,118.0-27,16,780.5 = 4,65,337.5

$$Block\ S.S. = \frac{1}{4}[(1{,}156)^2 + (1{,}133)^2 + (1{,}157)^2 + (1{,}134)^2 + (1{,}168)^2 + (1{,}201)^2 + (1{,}209)^2 + (1{,}166)^2] - C.F.$$

$$= \frac{1{,}08{,}72{,}492}{4} - 27{,}16{,}780.5 = 1{,}342.5$$

∴ Error S.S. = Total S.S. – S.S. due to Blocks – S.S. due to treatments

= 465337.5 – 1342.5 - 456857.5 = 7137.5

ANOVA Table

Source of Variation	***Degree of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Block	7	1,342.5	191.8	<1

Treatments	6	4,56,857.5	7,360.5	29.11
N	1	3,612.5	3,612.5	9.11
K	1	1,60,178.0	1,60,178.0	40.90
P	1	2,77,512.5	2,77.512.5	699.90
NK	1	392.0	392.0	0.98
NP	1	882.0	882.0	2.20
KP	1	14,280.5	14,280.5	36.01
Error	18	7,137.5	396.5	
Total	31	4,65,337.5		

Since calculated value of F for blocks is less than 2.59, the tabulated value of F for (7,31) d.f. at 5% 'level of significance', we fail to reject the null hypothesis.

H_o: Confounding is not effective

Hence, we conclude that confounding is not effective.

6.4.2 Confounding in 2^4 Factorial Experiment in 2^2 Blocks

In these 2^4 experiments we have 16 treatment combinations and these treatment combinations are allotted to the 4 blocks in a replicate. The total no. of confounding is 2^2-1 = 3 in which 2^2-1-2 = 1 will be generalised interaction. Suppose we select ABC and ABD confounding. Then,

ABCABD = $A^2 B^2$ CD = CD will also be confounded.

The 16 treatment combinations in a systematic order as follows:

(1), (a), (b), (ab), (c), (ac), (bc), (abc), (d), (ad), (bd), (cd), (abd), (acd), (bcd), (abcd).

Block-I	***Block-II***	***Block-III***	***Block-IV***
(1)	(a)	(b)	(c)
(ab)	(d)	(abc)	(abd)
(acd)	(cd)	(ad)	(ac)
(bcd)	(abcd)	(bd)	(bc)

6.4.3 Confounding in 2^n Factorial Experiment in 2^k Blocks per replicate

Let a 2^n experiment conducted in 2k blocks (k=2,3….) or $(2 \leq k \leq n)$ of equal size per replicate. Then:

1. The total no. of treatment combinations = 2^n

No. of combinations per block = $\frac{2^n}{2^k}$ = 2^{n-k}

Thus, we have 2^{n-k} treatment combinations in each block and these are assigned at random in each block. In each replicate there are 2^k blocks total, giving rise to $2^k - 1$ treatment contrast in the replicate.

2. Generalized Interaction: The interaction obtained on multiplying the symbols in two effects (interaction) together and equating the square of any letter equal to unity is called the General Interaction of the given effect.

Example: For any two effects X and Y the generalized interaction is given as:

X	***Y***	***Generalized Interaction***
A	BC	ABC
AB	CD	ABCD
ABC	ACD	ABC ACD = A^2BC^2D = BD
ABC	CD	ABC CD = ABC^2D = ABD

3. If a replicate is divided into 2^k blocks, then $2^k - 1$ effects are confounded, k of which are independent and remaining $2^k - 1 - k$ are their linear combinations or their Generalised interaction.
4. If the size of the block is 2^p, then apart from the control treatment, p of the treatment combinations is independent and the rest $2^p - 1 - p$ will be their generalization interaction or linear combinations.

If 2^n experiment is conducted in 2^k blocks (of equal size) in a replicate then the block size in each replicate is 2^{n-k}. Hence a part from the control treatment, n-k treatments will be independent and the remaining $(2^{n-k} - 1 - (n-k))$ treatments will be their linear combinations.

6.4.4 Confounding in more than two blocks

For 2^4 factorial experiment two blocks per replications are reasonable in experiment with a larger no. of factors, we, however use more blocks (greater than two) per replicate. Confounding with two groups one interaction is confounded in such cases. The Key Blocks contents are obtained from the solution of more than one homogenous equation simultaneously.

Example: The key block of size 2^3 in 2^5 can be obtained from the solutions of the equations.

$x_1 + x_2 + x_3 = 0$

$x_1 + x_4 + x_5 = 0$

These equations indicate that the interaction ABC and ADE are confounded simultaneously in the same replications. Any solutions of the above two homogenous equations are also a solution of the equation, which is obtained from a linear combination of the equations. This shows that the same interaction is also confounded. By adding the above equations, we get:

$x_2 + x_3 + x_4 + x_5 = 0$

No other equation is possible from their linear combination. Therefore, interaction BCDE is also confounded, which is the generalized interaction of ABC and ADE. They block can be obtained by first obtaining three independent solutions of the homogenous equations and then taking all their linear combinations.

Key Blocks

0	0	0	0	0
0	1	1	1	1
1	0	1	0	1
1	0	1	1	0
1	1	0	1	0
1	1	0	0	1
0	0	0	1	1
0	1	1	0	0

There are three more blocks in a replicate which are obtained from the solutions of the following three sets of the equations:

Set of equations	***IInd Block***	***IIIrd Block***	***IVth Block***
$x_1 + x_2 + x_3$	1	0	1
$x_1 + x_4 + x_5$	0	1	1

6.4.5 Comparison of Unconfounded and Completely Confounded 2^n Factorial Experiment

We know the information of effects contained in an experiment is the reciprocal of the variable of its estimate. In the case of unconfounded design, the replicate is itself a block, and, in this case, we denote the error variance by σ^2. In a completely confounded design, the replicate contains two blocks i.e., a block is a half replicate. Therefore, in this case we denote the error variance by $\sigma_{y_2}^2$ and it is expected that $\sigma_{y_2}^2 < \sigma^2$,since the smaller block will have greater control over errors as compare to the large block.

The variance of estimators of an effect (main or interactions) in a 2^n experiment in q replicated without confounding is $\sigma^2/q * 2^{(n-2)}$, whereas the variance of the estimator of each unconfounded effect in a 2^n experiment is completely confounded is $\sigma_{y_2}^2/q * 2^{(n-2)}$.

Thus, the information about each effect in an unconfounded design is $q * 2^{(n-2)}/\sigma^2$. Where as the information about each confounded effect in a completely con-founded design is $q * 2^{(n-2)}/\sigma_{(Y_2)}^2$ and since $\sigma_{y_2}^2<\sigma^2$,therefore the confounded design provides more information on unconfounded effect design but the confounded design provides no information or zero information has been completely confounded.

Note: When we are unsure which interactions are unimportant, we cannot sacrifice the entire interaction on that treatment combination. In such cases, we distribute the loss of information among more than one treatment combination and we shall get some information on each. This can be achieved by partial confounding.

6.5 Partial Confounding in Factorial Experiment

We have 4 interactions AB, AC, BC and ABC. We take 4 replications and two blocks of size 4 in each replication. The 8 treatment combinations are allotted to blocks of a replicate in such a way that the interaction AB is confounded in replicate I, AC in replicate II, BC in replicate III and ABC in replicate IV. The layout before randomization will be given as: -

Replicate-1 (AB confounded)

Block-1	*Block-2*
(1)	(a)
(ab)	(b)
(c)	(ac)
(abc)	(bc)

Replicate-2 (AC confounded)

Block-3	*Block-4*
(1)	(a)
(b)	(ab)
(ac)	(c)
(abc)	(bc)

Replicate-3 (BC confounded)

Block-5	*Block-6*
(1)	(b)
(a)	(ab)
(bc)	(c)
(abc)	(ac)

Replicate-4 (ABC confounded)

Block-7	*Block-8*
(1)	(abc)
(ab)	(a)
(ac)	(b)
(bc)	(c)

The block sum of squares is computed from the 8 blocks and grand totals. the sum of squares due to the main effect A, B and C (unconfounded effects) are computed using data from all 4 replications whereas the sum of squares due to any confounded interaction is obtained from those replicates where that particular interaction is not confounded.

The first two columns of ANOVA table in this case are given as below:

Source of Variation	***Degree of Freedom***
Blocks	7
A	1
B	1
C	1
AB	1
AC	1
BC	1
ABC	1
Error	17
Total	31

Example: Analyze the following 2^3 – Factorial experiment in blocks of 4 plots, involving three fertilizers N, P and K, each at two levels.

Replicate-I					***Replicate-II***					***Replicate-III***				
Block 1	np 101	npk 111	(1) 75	k 55	*Block 3*	(1) 125	npk 95	nk 80	p 100	*Block 5*	pk 75	nk 100	(1) 55	np 92
Block 2	p 88	n 90	pk 115	nk 75	*Block 4*	n 80	k 95	np 115	pk 90	*Block 6*	n 53	p 65	k 82	npk 76

Solution: Since each replicate has been divided into 2 blocks, one effect has been confounded in each replicate. Replicate 1 confounds NP, replicate *II* confounds NK and NPK has been confounded in replicate *III*.

Ho: The data is homogenous with respect to blocks and treatments.

Taking deviations from 87, we prepare the following Table to compute the total S.S. and S.S. for Blocks.

Calculations For Various S.S.

Treatment Combination	*Replicate I*		*Replicate II*		*Replicate III*		*Treatment Totals*
	Block 1	*Block 2*	*Block 3*	*Block 4*	*Block 5*	*Block 6*	
(1)	-12	-	38	-	-32	-	-6
n	-	3	-	-7	-	-34	-38
p	-	1	13	-	-	-22	-8
np	14	-	-	28	5-	-	47
k	-32	-	-	8	-	-5	-29
nk	-	-12	-7	-	13	-	-6
pk	-	28	-	3	-12	-	19
npk	24	-	8	-	-	-11	21
Block totals $(\boldsymbol{B_i})$	-6	20	52	32	-26	-72	G = 0
$\boldsymbol{B_i^2}$	36	400	2,704	1,024	676	5,184	$\sum B_i^2 = 10{,}024$

Correction Factor = $\frac{G^2}{3\times 8} = 0$

$R.S.S. = (-12)^2 + (38)^2 + \cdots + 8^2 + (-11)^2 = 8{,}658$

Total S.S. = R.S.S. - C.F. = 8,658

S.S. due to Blocks = $\sum_i \frac{B_i^2}{4} - C.F. = \frac{10{,}024}{4} = 2{,}506$

The S.S. due to interactions NP, NK and NPK are not estimable directly from the table of Yates' method, but they will be estimated indirectly.

Yates' Method for 2^3 Partially Confounded Experiment

Treatment Combination	*Total Yield*	*Yates' Operations*			*S.S.* = $\frac{[\,]^2}{8\times 3}$
		I	*II*	*Factorial Effects III*	
'1'	-6	-44	-5	0 = G	
n	-38	39	5	48 = [N]	$S_N^2 = 96.00$
p	-8	-35	23	158 = [P]	$S_P^2 = 1{,}040.17$
np	47	40	25	66 = [NP]	Not estimable
k	-29	-32	83	10 = [K]	$S_k^2 = 4.17$

nk	-6	55	75	2 = [NK]	Not estimable
pk	19	232	87	-8 = [PK]	$S^2_{PK} = 2.67$
npk	21	2	-21	-108 = [NPK]	Not estimable

Interaction, which is confounded in replicate 1, is estimated by:

$NP = \frac{1}{4}[(n-1)(p-1)(k+1)]$

Here the sign of '1' is positive. Hence, the adjusting factor (A.F.) for NP which is to be obtained from replicate 1 is given by:

A.F. for NP = (101+111+75+55) - (88+90+115+75) = 342-368 = -26

∴ Adjusted effect total for NP becomes: $[NP^*] = [NP] - (-26) = 66 + 26 = 92$

Similarly, A.F. for NK = 400-380 = 20

and, A.F. for NPK = 276-322 = -46 [Note that the sign of 1 in the estimate of NPK is -1.]

Hence, adjusted effect totals for NK and NPK are:

$[NK]^* = 2 - 20 = -18 \quad and\ [NPK]^* = -108 - (-46) = -62$

$S^2_{NP} = S.S.due\ to\ NP = \frac{1}{2\times 8}[NP^*]^2 = \frac{(92)^2}{16} = 529$

$S^2_{NK} = S.S.due\ to\ NK = \frac{1}{2\times 8}[NK^*]^2 = \frac{(-18)^2}{16} = 20.25$

$S^2_{NPK} = S.S.due\ to\ NPK = \frac{1}{2\times 8}[NPK^*]^2 = \frac{(-62)^2}{16} = 240.25$

$Treatment\ S.S. = S^2_N = S^2_P + S^2_K + S^2_{NP} + S^2_{NK} + S^2_{PK} + S^2_{NPK} = 1932.51$

∴ Error S.S. = T.S.S. – S.S. Blocks -S.S. Treatments

= 8,658.00-2,506-1,932.75 = 4,219.25

ANOVA Table for Partially Confounded 2^3 Experiment

Source of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio F***	
				$F_{Cal.}$	***$F_{Tab.}$***
Blocks	5	2,506.00	501.2	1.31	
Treatments	7	1,932.51	276.07	<1	$F_{0.05}(5,11) = 3.2$

N	1	96.00	96.00	<1	$F_{0.01}(5{,}11) = 5.32$
P	1	1,040.12	1,040.12	2.71	$F_{0.05}(1{,}11) = 4.84$
NP	1	529.00	529.00	0.38	$F_{0.01}(1{,}11) = 6.08$
K	1	4.17	4.17	<1	
NK	1	20.25	20.25	<1	
PK	1	2.67	2.67	<1	
NPK	1	240.25	240.25	<1	
Error	11	4,219.25	383.57		
Total	23	8,658			

From the above table, it can be concluded that the effect due to blocks, main effects due to factor N, P, and K or interactions are not significant.

6.5.1 Comparison of information about Unconfounded Effects and Confounded Effects in Partially Confounding Design

In partially confounding design the main effects A, B and C are not confounded in any replicates, so they are estimated from all 4 replicates and the experiment contains $8/\sigma^2_{(y_2)}$ information about each of the main effect but each interaction is confounded in one replicate and let unconfounded in three others. Thus, we can estimate the interaction from those replicates where it is not confounded.

Example: The interaction AB is confounded from replicate 1st; therefore, AB can be estimated using replicate 2,3, and 4 .so only three replicates control the information about the interaction AB and the amount of information is $\sigma/\sigma^2_{(y_2)}$,thus the relative information of each partially confounded interaction with respect to unconfounded main effect is $\left(6/\sigma^2_{(y_2)}\right)/\left(8/\sigma^2_{(y_2)}\right) = 3/4$ and this is the same as proportion of replicates given information about the confounded interaction.

Table for amount of information in different 2^3 experiments

Effects	***Amount of the information***		
	Unconfounded Design	***ABC completely confounded***	***AB, AC, BC & ABC partially confounded***
A	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$8/\sigma^2_{(y_2)}$

B	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$8/\sigma^2_{(y_2)}$
C	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$8/\sigma^2_{(y_2)}$
AB	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$8/\sigma^2_{(y_2)}$
AC	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$6/\sigma^2_{(y_2)}$
BC	$8/\sigma^2$	$8/\sigma^2_{(y_2)}$	$6/\sigma^2_{(y_2)}$
ABC	$8/\sigma^2$	*0*	$6/\sigma^2_{(y_2)}$

6.6 Self-Assessment Exercise

Question-1: Analyze the following 2^3 factorial design by determining the confounded treatment:

Replicate – 1

Block 1	1	nk	np	kp
	99	201	312	379
Block 2	npk	n	k	p
	408	98	260	306

Replicate – 2

Block 1	nk	np	1	kp
	308	352	100	412
Block 2	k	n	npk	p
	251	87	452	378

Replicate – 3

Block 1	np	nk	1	kp
	324	378	84	435
Block 2	n	npk	p	k
	135	456	378	272

Replicate – 4

Block 1	1	kp	nk	np
	99	201	312	379
Block 2	npk	p	k	n
	408	98	260	306

Question-2: A 2^3 experiment with factors a, b, c is to be conducted in 4 replicates consisting of two 4 plots blocks. Two experimenters conducted such 4- replicate experiments in two different farms – in one experiment ABC is totally confounded and in the other AB, AC, BC, and ABC are partially confounded. How will you make a combined analysis of this experiments?

6.7 Summary

This unit provides a brief overview of the confounding and partial confounding in the factorial experiments, focusing on the 2^3 and 2^4 factorial experiments and the concept of orthogonality.

6.8 References

- Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer-Verlag, New York.

- Gupta S.C., Kapoor V.K. (2007). *Fundamentals of Applied Statistics*, Sultan Chand & Sons.
- Gupta S.P. (2021). *Statistical Methods*, Sultan Chand & Sons.
- Gupta V.K. & Nigam A.K. Nigam (1978-79). *Handbook an analysis of Agriculture Experiment*, IASRI Publication.

6.9 Further Reading

- Cochran, W. G. and Cox, G. M. (1957). *Experimental Design*. John Wiley & Sons, New York.
- Cochran, W.G. and Cox, G.M. (2003). *Experimental Design*, Asia Publishing House
- Das, M. N. and Giri, N. S. (1986). *Design and Analysis of Experiments (2nd Edition)*. Wiley.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*. New Age International (P) Ltd. New Delhi.
- Montgomery, D.C. (2017). *Design and Analysis of Experiments (9th Edition)*. John Wiley & Sons.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments*, Wiley, New York.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 201N/ MASTAT – 201N Linear Models and Design of Experiment

Block: 3 Advance Theory of Design of Experiment

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Shambhavi Mishra **Writer**
Department of Statistics
University of Lucknow, Lucknow

Prof. Shruti **Editor**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Shruti **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 201N/ MASTAT – 201N **LINEAR MODEL & DESIGN OF EXPERIMENT**
©UPRTOU
First Edition: ***July 2023***
ISBN :

Printed By:

Block & Units Introduction

The ***Block - 3 – Advanced Theory of Design of Experiment*** is the last block of the said SLM, and it has three units.

Unit – 7 – BIBD and PBIBD dealt with Balanced Incomplete Block Design (BIBD), Partially Balanced Incomplete Block Design (PBIBD), construction of BIBD and PBIBD, association schemes and construction, resolvable and affine resolvable design.

Unit – 8 – Split and Strip Plot Design, comprises the Intra block and inter block analysis, Split Plot Design, Strip Plot Design.

In ***Unit – 9 – Other Advanced Design***, we have discussed the Dual and linked block design, Lattice Designs, Cross-over designs, optimal designs- optimal criteria, robust parameter design, response surface design – orthogonality, rotatability and blocking, weighing designs, mixture experiments

At the end of every unit the summary, self-assessment questions and further readings are given.

UNIT:7 BIBD & PBIBD

Structure

7.1 Introduction

In certain experiments using randomized block designs, we may not be able to run all the treatment combinations in each block. Situations like this usually occur because of shortages of experimental apparatus or facilities or the physical size of the block. For this type of problem, it is possible to use randomized block designs in which every treatment is not present in every block. These designs are known as randomized incomplete block designs. When all treatment comparisons are equally important, the treatment combinations used in each block should be selected in a balanced manner, so that any pair of treatments occur together the same number of times as any other pair. Thus, a Balanced Incomplete Block Design (BIBD) is an incomplete block design in which any two treatments appear together an equal number of times. Suppose that there are *a* treatment and that each block can hold exactly *k (k <a)* treatments. A balanced incomplete block design may be constructed by taking $\binom{a}{k}$ blocks and assigning a different combination of treatments to each block. Frequently, however, balance can be obtained with fewer than $\binom{a}{k}$ blocks.

Complete Block Design

When the block size is equal to the number of treatments used in the design or in other words when all the treatments appear in each block only once, the block design is said to be a *complete block design.*

Incomplete Block Design

When the block size is less than the number of treatments, *i.e.,* the block size is not equal to (but less than) the number of treatments, then we have an in complete block design. In other words, an incomplete block design is one in which the block size k is less than the number of treatments v, *i.e.,* $k < v$.

Equi-Replicate Design

An incomplete block design is said to be equi-replicate if all the treatments are replicated the same number of times.

Proper Design

An incomplete block design is said to be proper if the block size k is same for all the blocks.

Incidence Matrix

Suppose v treatments are applied in b blocks each of size k, where $k < v$, then we have an incomplete block design. Now consider a matrix $\boldsymbol{N}$ with b rows and v columns with elements as the number of observations receiving the different treatments in different blocks, *i.e.*

$\boldsymbol{N} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1v} \\ n_{21} & n_{22} & \cdots & n_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ n_{b1} & n_{b2} & \cdots & n_{bv} \end{bmatrix}$, where n_{ij} is the number of observations receiving i^{th} treatment in the j^{th} block. Obviously $n_{ij} = 1$ or 0 according to the i^{th} treatment is occurring in the j^{th} block or not.

Binary Design

An incomplete block design is said to be a binary design, if the elements of the incidence matrix of the design takes only two values namely 0 or 1.

Now let the number of times the i^{th} and l^{h} treatments occurring together in all the blocks be denoted by λ_{il}, *i.e.*, λ_{il} = the number of times the treatments *i* and *l* are appearing together in all the blocks. Then the parameters of an incomplete block design are *b, k, v, r* and λ_{il}, where *r* is the number of times each treatment is replicated or repeated.

7.2 Objectives

After going through this unit, you should be able to:

- Understand the principles of Balanced Incomplete Block Design (BIBD) and Partially Balanced Incomplete Block Design (PBIBD),
- Construct both BIBD and PBIBD,
- Know about resolvable design and affine resolvable design.

7.3 Balanced Incomplete Block Design (BIBD)

In a confounded design, we sacrifice some or total information on any of the treatment combination to maintain the homogeneity. Sometimes, we come across the situation where all the treatment (combination) has equal importance. Then, we cannot afford to sacrifice information on any of them by confounding. In that type of situations, we use Balanced Incomplete Block Designs (BIBD). This design was first proposed by Yates in 1936.

Here, incomplete blocks mean, a block which do not contain complete set of treatment. The use of incomplete blocks becomes necessary because as the no. of the treatments increases, the blocks size increases which results an increase the heterogeneity as well as the experimental error in a block. When the number of replications of all pairs of treatment in an incomplete block is the same than an important series of design known as Balanced Incomplete Block Design (BIBD). In this design all the treatment effects are estimated with equal precision.

A BIBD is an arrangement of ‘v’ treatments in ‘b’ blocks of size ‘k’ each such that each treatment is repeated ‘r’ times and the no. of times any pairs of treatments accrue together in all the blocks is ‘λ’. (v, b, k, r and λ are known) as the parameter of the BIBD.

7.3.1 Parameters of BIBD

BIBD has five parameters as given below:

v – Number of treatments in a replicate

r – Number of replicates for each treatment

b – Number of blocks (in r replicates)

k – Block size or number of plots in each block

λ – Number of blocks in which any pairs of treatments occur together or number of replications of each treatment pair.

7.3.2 Relationships among the Parameters of BIBD

The following are the relationship among the parameters of the BIBD:

(i) $bk = rv$

(ii) $r(k-1) = \lambda (v-1)$

(iii) $b \geq v$

(iv) $b \geq r + v - k$

Proof:

(i) The L.H.S. of bk = rv i.e., bk is the total no. of blocks in design and the R.H.S i.e., r.v gives the total no. of treatments which are to be used for the total no. of plots in the design consequently. L.H.S. = R.H.S.

(ii) L.H.S. of the equation r(k-1) = λ(v-1) is r (k-1). Since the block size is k, therefore any treatment Ti accruing in a block will form (k-1) pair with other treatments. Also since each treatment is repeated r times, therefore the treatment T_i will be occurring in r blocks and consequently Ti will be forming r (k-1) pairs with other treatments i.e., we have r (k-1) total no. of pairs a treatment t_i can form with other treatments.

The R.H.S. is equal to λ (v-1). Since the total no. of treatments is V, therefore any treatment Ti, can form (v-1) pairs with other treatments. But since any treatment T_i can form only λ pairs with any other treatment, hence the total no. of pairs which any particular treatment Ti can forms with the rest of the treatment λ (v-1) i.e., λ (v-1) is equal to the no. of pairs which any treatment t_i can form with the rest of the treatments. Hence r (k-1) = λ (v-1).

(iii) For proving the $b > v$ consider the incidence matrix N of BIBD.

$$\underline{N} = \begin{pmatrix} n_{11} & n_{12} & n_{ij} --- n_{ip} \\ n_{21} & n_{22} & n_{2j} --- n_{2b} \\ \vdots & \vdots & \vdots \;----\; \vdots \\ n_{i1} & n_{i2} & n_{ij} --- n_{ib} \\ \vdots & \vdots & \vdots ----\vdots \\ n_{v1} & n_{v2} & n_{vj} --- n_{vb} \end{pmatrix}$$

$$\underline{N}' = \begin{pmatrix} n_{11} & n_{12} & n_{ij} --- n_{ip} \\ n_{21} & n_{22} & n_{2j} --- n_{2b} \\ \vdots & \vdots & \vdots \;----\; \vdots \\ n_{i1} & n_{i2} & n_{ij} --- n_{ib} \\ \vdots & \vdots & \vdots ----\vdots \\ n_{v1} & n_{v2} & n_{vj} --- n_{vb} \end{pmatrix}$$

$$\underline{N}\,\underline{N}' = \begin{pmatrix} \sum_{j=1}^{b} n^2{}_{1j} & \sum_{j=1}^{b} n_{ij}n_{2j} & \dots & \sum_{j=1}^{b} n_{ij}n_{vj} \\ \sum_{j=1}^{b} n_{2j}n_{ij} & \sum_{j=1}^{b} n^2{}_{2j} & \dots & \sum_{j=1}^{b} n_{2j}n_{vj} \\ \vdots & \vdots & \dots & \vdots \\ \sum_{j=1}^{b} n_{vj}n_{ij} & \sum_{j=1}^{b} n_{vj}n_{2j} & \dots & \sum_{j=1}^{b} n_{vj}^2 \end{pmatrix}$$

$\because n_{ij}^2 = n_{ij}$ *then*

$$\underline{N}\,\underline{N}' = \begin{pmatrix} \sum_{j=1}^{b} n_{ij} & \sum_{j} n_{ij}n_{2j} & \dots & \sum_{j} n_{ij}n_{vj} \\ \sum_{j=1}^{b} n_{2j}n_{ij} & \sum_{j} n_{2j} & \dots & \sum_{j=1}^{b} n_{2j}n_{vj} \\ \vdots & \vdots & \dots & \vdots \\ \sum n_{vj}n_{ij} & \sum n_{vj}n_{2j} & \dots & \sum n_{vj} \end{pmatrix}_{v\times v}$$

$$= \begin{pmatrix} r & \lambda & \dots & \lambda \\ \lambda & r & \dots & \lambda \\ \vdots & \vdots & \dots & \vdots \\ \lambda & \lambda & \dots & r \end{pmatrix}; \begin{bmatrix} \because \sum_{j=1}^{b} n_{ij} = r \\ \sum_{j=1}^{b} n_{ij}\, n_{il} = \lambda \end{bmatrix}$$

$$= \begin{pmatrix} r-\lambda & 0 & \dots & 0 \\ 0 & r-\lambda & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & r-\lambda \end{pmatrix} + \begin{pmatrix} \lambda & \lambda & \dots & \lambda \\ \lambda & r & \dots & \lambda \\ \vdots & \vdots & \dots & \vdots \\ \lambda & \lambda & \dots & \lambda \end{pmatrix}$$

$$= I_v(r-\lambda) + J_{vv}\lambda$$

$$\textit{where}\quad I_v = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{v\times v} \quad \textit{and}\ J_{vv} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

J_{vv} be square matrix of order $v \times v$ now consider the determinant of $\underline{N}\,\underline{N}'$

$$\left|\underline{N}\,\underline{N}'\right| = \begin{vmatrix} r & \lambda & \cdots & \lambda \\ \lambda & r & \cdots & \lambda \\ \vdots & \vdots & \cdots & \vdots \\ \lambda & \lambda & \cdots & r \end{vmatrix}$$

Now adding $(v-1)\lambda$ with the first row

$$\left|\underline{N}\,\underline{N}'\right| = \begin{vmatrix} r+(v-1)\lambda & r+(v-1)\lambda & \cdots & r+(v-1)\lambda \\ \lambda & r & \cdots & \lambda \\ \vdots & \vdots & \cdots & \vdots \\ \lambda & \lambda & \cdots & r \end{vmatrix}$$

$$\left|\underline{N}\,\underline{N}'\right| = r+(v-1)\begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda & r & \cdots & \lambda \\ \vdots & \vdots & \cdots & \vdots \\ \lambda & \lambda & \cdots & r \end{vmatrix}$$

Now multiplying first row by and subtracting it for the remaining rows.

$$\left|\underline{N}\,\underline{N}'\right| = r+(v-1)\begin{vmatrix} 1 & 1 & 1 & --- & 1 \\ 0 & r-\lambda & 0 & ------- & 0 \\ 0 & 0 & r-\lambda & ---- & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ---- & r-\lambda \end{vmatrix}$$

$$= \{r-(v-r)\lambda\}(r-\lambda)^{v-1}$$

But from relation (ii) $(v-r)\lambda = r(k-1)$

$\therefore\ \left|\underline{N}\,\underline{N}'\right|\{r + r(k-1)\}(r-\lambda)^{v-1}$

$= r\ k\ (r-\lambda)^{v-1} \neq 0$

$Since\left|\underline{N}\,\underline{N}'\right| \neq 0, \therefore rank\ \left(\underline{N}\,\underline{N}'\right) = v = r\left(\underline{N}\right).$

$(iv)\ To\ prove\ b\ \geq r+v-k.$

Since we have $b\ \geq r$ and bk = rv

$\therefore k \geq r or r \geq k \Longrightarrow r-k \geq 0$ (1)

Also, since a BIBD is an incomplete block design, hence v>k⟹ v-k > 0 (2)

From (i) & (ii) we have:

$(r-k)\ (v-k) \geq 0.$

$$or\ \left(\frac{r}{k}-1\right)\left(\frac{v}{k}-1\right) \geq 0$$

$$or\ \frac{r.v}{k^2}-\frac{r}{k}-\frac{v}{k}+1 \geq 0$$

$$or\ \frac{b.k}{k^2}-\frac{r}{k}-\frac{v}{k}+1 \geq 0;\ (since\ bk = rv)$$

$$or\ b-r-v+k\ \geq 0$$

$$\Rightarrow b \geq r+v-k$$

Proved.

Note: In case of BIBD if b-v the BIBD is known as the *Symmetric BIBD*

***Cor:* -** In a symmetric BIBD (i.e., b=v) and if v is even then (r-v) must be a perfect square.

***Proof:* -** Consider the incidence matrix

$$\underline{N} = \begin{pmatrix} n_{11} & \dots & n_{iv} \\ \vdots & \vdots & \vdots \\ n_{v1} & \dots & n_{vv} \end{pmatrix}; since\ b = v$$

$$\underline{N}\,\underline{N}' = \begin{pmatrix} r & \lambda & \dots & \lambda \\ \lambda & r & \dots & \lambda \\ \vdots & \vdots & \dots & \vdots \\ \lambda & \lambda & \dots & r \end{pmatrix}$$

$$\left|\underline{N}\,\underline{N}'\right| = \{r+(v-1)\lambda\}(r-\lambda)^{v-1}$$

$$= \{r+r(k-1)\}(r-\lambda)^{v-1}$$

$$= rk\ (r-\lambda)^{v-1} = r^2(r-\lambda)^{v-1}$$

$$(Since\ bk = r.v\ or\ vk = rv\ \Rightarrow k = r)$$

$$or\ \left|\underline{N}\,\underline{N}'\right| = \left|\underline{N}\right|\left|\underline{N}'\right|$$

$$\left|\underline{N}'\right| = +r(r-\lambda)^{(v-1)/2} \qquad (1)$$

Since $\underline{N}$ is a matrix of 0 and 1, hence determinent $\underline{N}$ will be an integer from eq^n (I)

But if v is even then determinant $\underline{N}$ will be an integer only when $(v-\lambda)$ is a perfect.

Proved.

7.4 Balanced Design

A balanced design is a design in which all the elementary contrast are estimated with the same precision i.e. The var. of each elementary contrast is the same.

Note:- In the case of BIBD we have seen that $v(\hat{\alpha_i} - \hat{\alpha_j}) = \frac{2k}{\lambda v}\sigma^2; which\ is\ indeep\ of\ i\ \&\ j,$ which means that $v(\hat{\alpha_i} - \hat{\alpha_j})$ is same for all i and j it is in this sence we say that this is a Balanced Incomplete Block design.

Theorem: In symmetrical BIBD the no of treatments common between any two block is λ.

OR

In symmetrical BIBD every block different from the first 2 block has exactly λ treatments common with the first block.

Proof: Let ai be the no of treatments common to the first and i^{th} block, then since each of the k treatment of first block occurs (r-1) times in the other blocks and hence

$$\sum_{i=2}^{b} ai = k(r-1) \qquad (1)$$

Again, since each of the $^kC_2$ pairs of the treatments of the first block occurs (λ-1) times in the remaining blocks.

$\sum_{i=2}^{b} {}^{ai}C_2 = (\lambda\text{-}1)\ ({}^kC_2)$

$\sum_{i=2}^{b} \frac{ai(ai-1)}{2} = \frac{(\lambda-1)k(k-1)}{2}$

$\sum_{i=2}^{b} ai^2 - \sum_{i=2}^{b} ai = (\lambda-1)k(k-1)$

$\sum_{i=2}^{b} ai^2 = (\lambda-1)k(k-1) + \sum_{i=2}^{b} ai \qquad (2)$

Let us consider:

$\sum_{i=2}^{b}(ai-\lambda)^2 = \sum_{i=2}^{b} ai^2 - 2\lambda \sum_{i=2}^{b} ai + \lambda^2 \sum_{i=2}^{b} 1$

$= k(\lambda-1)(k-1) + \sum_i ai - 2\lambda \sum_i ai + (b-1)\lambda^2$ using (1)

$= k(\lambda-1)(k-1) - (2\lambda-1) \sum_i ai + (b-1)\lambda^2$

$= k(\lambda-1)(k-1) - (2\lambda-1).\ K(r-1) + (b-1)\lambda^2$

$= k(k-1)\ \{\lambda-1-2\lambda+1\} + (b-1)\lambda^2$

$= -\lambda k(k-1) + (b-1)\lambda^2$

$= -k(k-1)\lambda + (t-1)\lambda^2$ {b=t for symmetric BIBD}

$= -k(k-1)\lambda + r(k-1)\lambda$ {$\lambda(t-1) = r(k-1)$}

$= -k(k-1)\lambda + k(k-1)\lambda$ {k=r for symmetric BIBD}

$= 0$

$\sum_{i=2}^{b}(ai-\lambda)^2 = 0$

$ai = \lambda$

Therefore, for symmetrical BIBD the no of treatment common between any two blocks is equal to no of pairs i.e., λ.

7.5 Analysis of BIBD

There are two types of analysis as follows:

7.5.1 Intra-Block Analysis of BIBD

Observation Table

Treatments	***Blocks***					
	1	***2***	**------**	***j***	**----**	***b***
1	$n_{11}y_{11}$	$n_{12}y_{12}$	……	$n_{1j}y_{1j}$		$n_{1b}y_{1b}$
2	$n_{21}y_{21}$	$n_{22}y_{22}$	……	$n_{2j}y_{2j}$		$n_{2b}y_{2b}$
: ***:***	: :	: :	: :	: :	: :	: :
i	$n_{i1}y_{i1}$	$n_{i2}y_{i2}$	……	$n_{ij}y_{ij}$		$n_{ib}y_{ib}$
: ***:***	: :	: :	: :	: :	: :	: :
v	$n_{v1}y_{v1}$	$n_{v2}y_{v2}$	……	$n_{vj}y_{vj}$		$n_{vb}y_{vb}$

Total	$\sum_{i=1}^{v} n_{ij}y_{ij} = T_{o1}$	$\sum_{i=1}^{v} n_{i2}y_{i2} = T_{o2}$	……	$\sum_{i=1}^{v} n_{ij}y_{ij} = T_{oj}$		$\sum_{i=1}^{v} n_{ib}y_{ib} = T_{ob}$

$\text{Totals} = \sum_{j=1}^{b} n_{ij}y_{ij} = T_{1o}, \sum_{j=1}^{b} n_{2j}y_{2j} = T_{2o} . - - - T_{io} - - - T_{vo}$

$\text{Agreegate Total} = \sum_{i=1}^{v} T_{io} = \sum_{i=1}^{v} T_{oj} = T_{00} = \sum_i \sum_j n_{ij}y_{ij}$

Mathematical Model:

$$n_{ij}y_{ij} = n_{ij}\left(\mu + \alpha_i + \beta_j + e_{ij}\right);\ j = 1, - - - b; i = 1;\ - - v$$

Assumption:

$$e_{ij}^{'}s\ are\ iid \sim N(0, \sigma^2)$$

Side condition = $\sum_i \alpha_i = \sum_j \beta_j = 0$

Unrestricted Residual Sum of Squares (RSS):

$SSE = \sum_i \sum_j e_{ij}^2 = SSE = \sum_i \sum_j n_{ij}^2 \left(y_{ij} - \mu - \alpha_i - \beta_j\right)^2$ $(since\ n_{ij}^2 = n_{ij})$

Differentiating this with respect to $\mu - \alpha_i - \beta_j$ and equating them individually to zero.

$\frac{\gamma SSE}{\partial \mu} = 0 \Rightarrow \sum_i \sum_j n_{ij}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$

$= \sum_i \sum_j n_{ij}\, y_{ij} = \sum_i \sum_j n_{ij}\, \mu + r\sum_i \alpha_i + k\sum_j \beta_j$

$(Since\ \sum_j n_{ij} = r\ and\ \sum_i n_{ij} = k)$

$or\ T_{..} = bk\mu = \hat{\mu} = \frac{T_{..}}{bk} = \bar{y}_{..}$ (i)

$\frac{\gamma SSE}{\partial \mu} = 0 = \sum_{j=1}^{b} n_{ij}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)$

$\sum_j n_{ij}\, y_{ij} = \sum_j n_{ij}\ \mu + \sum_j \alpha_i + \sum_j \beta_j\, n_{ij}$

$T_{io} = r\mu + r\alpha_i + \sum_j n_{ij}\, \beta_j$ (ii)

$\frac{\gamma SSE}{\partial \beta_j} = 0 = \sum_j^{b} n_{ij}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)$

$\sum_i n_{ij}\, y_{ij} = \sum_i n_{ij}\ \mu + \sum_i \alpha_i\, n_{ij} + \sum_i n_{ij}\, \beta_j$

$T_{ij} = k\mu + \sum_j n_{ij}\, \alpha_i + k\beta_j$ (iii)

From (iii) the estimate of β_j is $\beta_j^{\wedge} = \frac{1}{k}\left(T_{oj} - k\mu - \sum_{i=1}^{v} n_{ij}\alpha_i\right)$ substituting this equation (ii) we get:

$$T_{io} = r\mu + r\alpha_i + \sum_j n_{ij}\ \frac{1}{k}\left(T_{oj} - k\mu - \sum_{h=1}^{v} n_{ij}\alpha_i\right)$$

$$= \ r\mu + r\alpha_i + \frac{1}{k}\sum_j n_{ij}T_{oj} - \sum_j n_{ij}\,\mu - \frac{1}{k}\sum_j n_i j\ \sum_{l=1}^{v} n_{ij}\alpha_h$$

$$= r\mu + r\mu + \frac{1}{k}\sum_{j=1}^{b} n_{ij}\, T_{oj} - r\mu - \frac{1}{k}\left\{\sum_{j=1}^{b} n_{ij}^2\ \alpha_i + \sum_{j=1}^{b}\sum_{h(\neq l)=1}^{v} n_{ij} h_{\alpha\lambda}\right\}$$

$$= r\alpha_i + \frac{1}{k}\sum_{j=1}^{b} n_{ij}\, T_o j - \frac{1}{k}\left\{\left\{\sum_{j=1}^{b} n_{ij}^2\alpha_i + \sum_{j=1}^{b}\sum_{h(\neq l)=1}^{v} n_{ij}n_{hj}\alpha h\right\}\right.$$

$$or\left(T_{io} - \frac{1}{k}\sum_{j=1}^{b} n_i j\ \ T_o j\right) or\left(T_{io} - \frac{1}{k}\sum_{j=1}^{b} n_i j\ \ T_o j\right) = r\alpha_i - \frac{1}{k}\sum_{j=1}^{b}\lambda_{ij}\,\alpha_i - \frac{1}{k}\sum_{j=1}^{b}\sum_{h\neq i=1}^{v} n_{ij}\, n_{hj}\,\alpha h$$

$$or\ Q_i - r\alpha_i + \frac{1}{k}\sum_{j=1}^{b} n_{ij}\,\alpha_i + \frac{1}{k}\sum_{h\neq i=1}^{v}\left\{\sum_{j=1}^{b} n_{ij}\, n_{hj}\right\} = 0$$

Where,

$$Q_i = T_{io} - \frac{1}{k}\sum_{j=1}^{b} n_{ij}n_{gj} = \lambda$$

$$\therefore\ Q_i - r\alpha i + \frac{1}{k} r\alpha i + \frac{1}{k}\lambda\sum_{h(\neq i)=1}^{v}\alpha_h\ = 0$$

$$or\ Q_i - r\left(1 - \frac{1}{k}\right)\alpha_i + \frac{\lambda}{k}\left[\sum_{h=1}^{v}\alpha_h - \alpha_i\right] = 0$$

$$or\ Q_i - \frac{r(k-1)}{k}\alpha_i + \frac{\lambda}{k}[0 - \alpha_i] = 0\ (since\ \sum_{h=1}^{v}\alpha_h\ = 0)$$

$$or\ Q_i - \frac{r(k-1)+\lambda}{k}\ \alpha_i\ = 0$$

$$or\ Q_i - \frac{\lambda(v-1)}{k}\alpha_i\ = 0\ \left(Since\ r(k-1) = \lambda(v-1)\right)$$

$$or\ Q_i - \frac{\lambda v}{k}\alpha_i\ = 0$$

$$or\ \widehat{\alpha_\iota} = \frac{k}{\lambda v}\ Q_i$$

$$Therefore\ SSE\ = \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}\left(\mathcal{Y}_{ij} - \hat{\mu} - \widehat{\alpha_\iota} - \widehat{\beta_J}\right)^2$$

$$= \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}\ y_{ij}\left(\mathcal{Y}_{ij} - \hat{\mu} - \widehat{\alpha_\iota} - \widehat{\beta_J}\right)$$

$(since\ other\ terms\ will\ be\ zero\ because\ of\ normal\ equation)$

$$= \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}\, y_{ij}^2 - \hat{\mu}\ \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij} - \sum_{i-1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij} - \widehat{\alpha_\iota} -$$

$$\sum_{i-1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij} - \widehat{\beta_J}$$

$$= \sum_i\sum_j n_{ij}y_{ij}^2 - \hat{\mu}\, T_{oo} - \sum_j T_{io}\,\widehat{\alpha_\iota} - \sum_{i-1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij}\frac{1}{k}\big(T_{oj} - k\,\hat{\mu} -$$

$$\sum_{h=1}^{v} n_{nj}\alpha_h\big)$$

$$= \sum_i\sum_j n_{ij}y_{ij}^2 - \hat{\mu}\, T_{oo} - \sum_{i=1}^{v} T_{io}\,\widehat{\alpha_\iota} - \frac{1}{k}\sum_{j=1}^{b} T_{oj}^2 + \hat{\mu}\, T_{oo} +$$

$$\frac{1}{k}\sum_j T_{oj}\ \sum_{i=1}^{v} n_{ij}\,\alpha_i + \frac{1}{k}$$

$$= \sum_v\sum_j n_{ij}\ \mathcal{Y}_{ij}^2 - \frac{1}{k}\sum_j T_{oj}^2 - \sum_i\left(T_{io} - \frac{1}{k}\sum_i n_{ij}T_{oj}\right)\widehat{\alpha_\iota}$$

$This\ will\ have\ bk - b - (v-1) = bk - b - v + 1\ degrees\ of\ freedom.$

$$SSE^* = \sum_i\sum_j n_{ij}\left(y_{ij} - \mu - \beta_j\right)^2$$

$$= \sum_i\sum_j n_{ij}\left(y_{ij} - \mu - \beta_j\right)^2$$

$\#(since\ wo\ make\ a\ assumption\ \sum \alpha_i = 0)$

Differentiating this with respect to μ and β_j and equating them individually to zero.

$$\frac{\partial SSE^*}{\partial\mu} = 0 = \sum_i\sum_j n_{ij}\left(y_{ij} - \mu - \beta_j\right) = 0 \qquad \text{(iv)}$$

$$\frac{\partial SSE^*}{\partial\mu} = 0 = \sum_i n_{ij}\left(y_{ij} - \mu - \beta_j\right) = 0 \qquad \text{(v)}$$

From equation (iv)we can write $\hat{\mu} = \frac{T_{oo}}{bk}$ and from (v) we have $T_{oj} = k\big(\mu + \beta_j\big)\ or\ \hat{\mu} + \beta_j = \frac{T_{oj}}{k}$

$$\therefore SSE^* = \sum_i\sum_j n_{ij}\left(y_{ij} - \hat{\mu} - \beta_j\right)^2$$

$$= \sum_i\sum_j n_{ij}\mathcal{Y}_{ij}\left(y_{ij} - \hat{\mu} - \beta_j\right)$$

(Other terms being zero of normal equation (iv) & (v))

$$= \sum_i\sum_j n_{ij}\ \mathcal{Y}_{ij}^2 - \sum_i\sum_j n_{ij}\mathcal{Y}_{ij}\frac{T_{oj}}{k}$$

$$= \sum_i \sum_j n_{ij} \; \mathcal{Y}_{ij}^2 - \sum_j \frac{T_{oj}^2}{k} \; ; \; \text{this will have } bk - b \text{ d.f.}$$

$$\therefore \text{Sum of Square due to Treatment } (adj) = SSE^* - SSE \ (adjested)$$

$$= \sum_i \sum_j n_{ij} \; \mathcal{Y}_{ij}^2 - \sum_j \frac{T_{oj}^2}{k} - = \sum_i \sum_j n_{ij} \; \mathcal{Y}_{ij}^2 - \frac{1}{k} \sum_j T_{oj}^2 + \sum_j Q_i \, \hat{\alpha}_i$$

$$= \sum_j Q_i \alpha_i$$

It will have (v-1) d.f.

ANOVA Table for Intra Block Analysis of BIBD

Sources of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Blocks (unadjusted)	b-1	$SSB = \frac{1}{k} \sum_j T_{oj}^2 - CF$	MSB	
Treatments (adj)	v-1	$SST = \sum_j Q_i \alpha_i$	MST (adj)	$F = \frac{MST}{MSE}$
Error	bk-b-v+1	$SSE = \sum_i \sum_j n_{ij} \, y_{ij}^2 - \frac{1}{k} \sum_j T_{oj}^2 - \sum_j Q_i \alpha_i$	MSE	
Total	bk-j	$TSS = \sum_i \sum_j n_{ij} \; \mathcal{Y}_{ij}^2 - CF$		

***Note:* -** If b = v then $\underline{N}\,\underline{N}' = \underline{N}'\,\underline{N}$ and λ will be equal to the no. of treatments between any two blocks.

$$SST(adj) = \sum_j Q_i \alpha_i \quad \text{where } Q_i = T_{io} - \frac{1}{k} \sum_j n_{ij} \, T_{oj} \;\; \text{and } \hat{\alpha}_i = \frac{k}{\lambda v} \; Q_i$$

$$\# \; V = (T_{io}) v\left(\sum_i \sum_j n_{ij} \mathcal{Y}_{ij}\right) = v\left(\sum_i n_{ij} \, e_{ij}\right)$$

$$= \sum_j n_{ij} \;\; v\left(e_{ij}\right) = \sum_j n_{ij} \, \sigma^2 = r \, \sigma^2$$

$$v\left(T_{oj}\right) = v\left(\sum_j n_{ij} \, y_{ij}\right) = \sum_j n_{ij} \;\; v\left(C_{ij}\right)$$

Variance of the Difference Between Two Treatment Effects:

If α_i and α_j are the i[th] and j[th] treatment effects then:

$\alpha_i = \frac{k}{\lambda v} Q_i, \ \alpha_j = \frac{k}{\lambda v} \ Q_j$

$Therefore$:

$v\left(\alpha_i - \alpha_j\right) = v\left(\frac{k}{\lambda v} Q_i - \frac{k}{\lambda v} Q_j\right)$

$= \frac{k^2}{\lambda^2 v^2} var\left(Q_i - Q_j\right)$

$= \frac{k^2}{\lambda^2 v^2} var\left[v(Q_i) - v\left(Q_j\right) - 2\left(v\left(Q_i - Q_j\right)\right)\right]$

$Now\ Q_i = T_{io} - \frac{1}{k} \sum_j n_{ij} \ T_{oj}$

$or\ k\ Q_i = k\ T_{io} - \sum_j n_{ij} \ T_{oj}$

$= (k-1)\ T_{io} - \left(\sum_j n_{ij} \ T_{oj} - T_{io}\right)$

The expression $\sum_j n_{ij} \ T_{oj} - T_{io}$ is now the sum of rk-r = r(k-1) observations or treatments other than the ith treatment. Therefore,

$v(kQ_i) = k^2 v(Q_i) = (k-1)^2\ v\ (T_{io}) + v\left(\sum_j n_{ij} \ T_{oj} - T_{io}\right)$

Since covariance terms will be zero.

$= (k-1)^2 \ r\sigma^2 + r\ (k-1)\ \sigma^2$

$= r(k-1)[k-1+1]\ \sigma^2$

$= rk\ (k-1)\sigma^2$

Therefore $v(Q_i) = \frac{rk\ (k-1)}{k^2}\ \sigma^2$

$= \frac{r\ (k-1)}{k}\ \sigma^2;\ \forall i$

To obtain covariance between Q_i and Q_j we have $\sum_{i=1}^{v} Q_i = 0$, therfore we should have $V(\sum_{i=1}^{v} Q_i) = 0$.

$$or\ we\ have\ \sum_{i=1}^{v} v(Q_i) + \sum_{i \neq j} cov\left(Q_i - Q_j\right) = 0$$

$$\sum_{i=1}^{v} \frac{r(k-1)}{k}\sigma^2 + V\,(v-1)\,cov\,(Q_i - Q_j) = 0$$

$$or\ cov\,(Q_i - Q_j) = -\frac{r\,v(k-1)}{k\,v(v-1)}\,\sigma^2$$

$$= -\frac{r\,(k-1)}{k\,(v-1)}\,\sigma^2$$

$$= -\frac{\lambda\,(k-1)}{k\,(v-1)}\,\sigma^2$$

$$= -\frac{\lambda}{k}\,\sigma^2$$

$$where\ \lambda = \frac{r\,(k-1)}{(v-1)}$$

$$\therefore V(\hat{\alpha}_i - \hat{\alpha}_j) = \frac{k^2}{\lambda^2 v^2}\left[\frac{r(k-1)}{k}\sigma^2 + \frac{r(k-1)}{k}\sigma^2 + \frac{2\lambda}{k}\,\sigma^2\right]$$

$$= \frac{k}{\lambda^2 v^2}\sigma^2\,[2r\,(k-1)\sigma^2 + 2\lambda\,\sigma^2]$$

$$= \frac{2k}{\lambda^2 v^2}\sigma^2\,[r\,(k-1) + \lambda]$$

$(use\ r(k-1) = \lambda(v-1))$

$$= \frac{2k}{\lambda^2 v^2}\,[\lambda(v-1) + \lambda]\sigma^2$$

$$= \frac{2k}{\lambda v}\sigma^2; which\ is\ independent\ of\ i\ and\ j.$$

(This property is known as ***Balancing Property***)

7.5.1.1 Efficiency of BIBD Relative of Randomized Block Design (RBD)

The efficiency of a design is defined as a ratio E= V_r/v (V_r/v) Where V is the Variance of the estimated intra block treatment elementary contrast for design using the same no. of experimental units, V and V_r be incomplete on the assumption that the intra block error variance is same in both cases.

$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \bar{y}_{io} = \mu + \alpha_i + c_{io}$

$\bar{y}_{oo} = \mu + \bar{c}_{00}$

$\hat{\alpha}_i = \bar{y}_{io} - \bar{y}_{oo}$

$v(\alpha_i) = V(\bar{y}_{io} - \bar{y}_{oo})$

$= v(\alpha_i\ \bar{e}_{io} - \bar{e}_{oo})$

$= v(\bar{e}_{io}) + v(\bar{e}_{oo}) - 2(\bar{e}_{io} - \bar{e}_{oo})$

In R.B.D. $v(\hat{\alpha}_i - \hat{\alpha}_j) = v(\bar{y}_{io} - \bar{y}_{oo} - \bar{y}_{jo} - \bar{y}_{oo})$

$= v(\mu + \alpha_i + \bar{e}_{io} - \mu - \alpha_j - \bar{e}_{jo})$

$= v(\bar{e}_{io}) + v(\bar{e}_{jo})$

(The covariance term will be zero)

$= \frac{\sigma^2}{r} + \frac{\sigma^2}{r} = \frac{2\sigma^2}{r}$

Hence $v = var(\hat{\alpha}_i - \hat{\alpha}_j) = \frac{2k}{\lambda v}\sigma^2\ and\ V_k = v(\hat{\alpha}_i - \hat{\alpha}_j) = \frac{2\sigma^2}{r}$

Then:

$E = \frac{V_k}{V} = \frac{2\sigma^2}{r} \times \frac{\lambda v}{2k\sigma^2} = \frac{\lambda v}{rk}$

Also we have $r(k-1) - \lambda(v-1)$

$\therefore\ \frac{\lambda}{r} = \frac{(k-1)}{(v-1)}$

$\therefore\ E = \frac{(k-1)v}{k(v-1)} = \frac{\frac{1}{k}(k-1)}{\frac{1}{v}(v-1)} = \frac{1-\frac{1}{k}}{1-\frac{1}{v}}$

In a BIBD since v >k,

$\frac{1}{v} < \frac{1}{k}$

$\therefore\ 1 - \frac{1}{v} > 1 - \frac{1}{k}$

Therefore, $E = \frac{1-\frac{1}{k}}{1-\frac{1}{v}} < 1$

Hence BIBD is less efficient than RBD.

7.5.1.2 C-Matrix of a BIBD

C matrix of an incomplete block design is given by: $\underline{c} = \left(\left(c_{ij}\right)\right); i,\ j, = 1, - - - - v$

$Where\ C_{ii} = n_{ii} - \frac{\sum_j n_{ij}^2}{n_{ej}}\ \ r - \frac{r}{k} and$

$\underset{i \neq l}{\textbf{Cil}} = \sum_j \frac{\boldsymbol{n_{ij}}}{\boldsymbol{n_{oj}}} = \frac{\boldsymbol{-\lambda_{il}}}{\boldsymbol{k}}$

For a BIBD, $C_{ii} = n_{io} - \sum_j \frac{\boldsymbol{n_{ij}}}{\boldsymbol{n_{oj}}}$

$$= r - \sum_j \frac{\boldsymbol{n_{ij}}}{\mathrm{k}} = \mathrm{r} - \frac{\mathrm{r}}{\mathrm{k}}$$

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

$and = \sum_j \frac{n_{ij}\, n_{ij}}{\mathrm{k}} = \frac{-\lambda}{k}$

Hence the C- matrix is given by:

$$\underline{C} = \begin{pmatrix} \frac{r(k-1)}{k} & \frac{-\lambda}{k} & \frac{-\lambda}{k} - - - - - \frac{-\lambda}{k} \\ \frac{-\lambda}{k} & \frac{r(k-1)}{k} & \frac{-\lambda}{k} - - - - - \frac{-\lambda}{k} \\ \vdots & \vdots & \vdots \qquad\qquad \vdots \\ \frac{-\lambda}{k} - - - - - - - - - - - - - \frac{r(k-1)}{k} \end{pmatrix}$$

$$= \begin{pmatrix} r & 0 & - - - - 0 \\ 0 & r & - - - - - 0 \\ \vdots & \vdots & - - - - - \vdots \\ 0 & 0 & r \end{pmatrix} - \begin{pmatrix} \frac{r}{k} & \frac{\lambda}{k} & & - - - - \frac{\lambda}{k} \\ \frac{\lambda}{k} & \frac{r}{k} & & - - - - \frac{\lambda}{k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\lambda}{k} - & \frac{\lambda}{k} - - - - - - - - & & \frac{r}{k} \end{pmatrix}$$

$$= r\begin{pmatrix} 1 & 0 - - - & 0 \\ 0 & 1 - - - & 0 \\ \vdots & ; - - - & \vdots \\ 0 & 0 - - - - & 1 \end{pmatrix} - \frac{1}{k}\begin{pmatrix} r & \lambda & - - - - \lambda \\ \lambda & r & - - - - - \lambda \\ \vdots & \vdots & - - - - - \vdots \\ \lambda & \lambda & r \end{pmatrix}$$

$$= r\underline{I_v} - \frac{1}{k}\,\underline{N}\,\underline{N'}$$

7.5.2 Inter-Block Analysis of BIBD

Inter block Analysis of BIBD is an analysis with recovery of inter block information. In this case mathematical model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + C_{ij};\ \ i = 1 - - - v, j = 1 - - - b$$

Where,

μ be the general mean α_i is additive effect due is i[th] treatment, e_{ij} are independent normal vitiates with mean zero and variance $\sigma_e^2,\ i.e.\, e_{ij}\ \sim N(0,\sigma_e^2)$

β_j be the additive effect due to j[th] block and it is normally independently distributed with mean zero and variance $\sigma_b^2\ \ i.e.\, \beta_j^{'} s \sim N(0,\sigma_b^2)$

Also $\beta_j^{'} s$ are independent of $e_{ij}^{'} s$.

Now equation (i) can be written as:

$$n_{ij}\, y_{ij} = n_{ij}\, \left(\mu + \alpha_i + \beta_j + e_{ij}\right) \qquad \text{(ii)}$$

Summing (ii)over i,

$$\sum_j n_{ij}\ y_{ij} = \sum_i n_{ij}\, \mu + \sum_i \alpha_i n_{ij} + \sum_i \beta_j n_{ij} + \sum_j n_{ij}\, e_{ij}$$

Or,

$$T_{oj} = k\mu + \sum_j n_{ij}\, \widehat{\alpha_i} + f_j, where\ f_j = k\widehat{\beta_j} + \sum_i n_{ij}\, e_{ij}\ \ and$$

$$f_i \sim N(0,\sigma_p{}^2)\ where\ \sigma_f{}^2 - v\left(f_j\right) = k^2 v\left(\beta_j\right) + \sum_{i=1}^{v} n_{ij}^2\, v(c_{ij})$$

$$= k^2\ \sigma_p{}^2 + \sum_{i=1}^{v} n_{ij}^2\ \sigma_e{}^2$$

$$= k^2 \sigma_b{}^2 + \sigma_e{}^2\ \sum_i n_{ij}\,;\ \left(Since\ n_{ij}^2 = n_{ij} = 1\ or\ 0\right)$$

$$= k^2 \sigma_b{}^2 + k\ \sigma_e{}^2$$

$$= k(\sigma_e{}^2 + k\ \sigma_b{}^2)$$

Hence the Residual Sum of Square (RSS) is:

$$\sum_j^b f_j{}^2 = \sum_{j=1}^{b} \left(T_{oj} - k\mu - \sum_{i=1} h_{ij}\, \widehat{\alpha_i}\right)^2$$

Differentiating this w.r.t μ andα_i and equating to zero.

$$\sum_j \left(T_{oj} - k\mu - \sum_{i=1}^{v} n_{ij}\alpha_i\right) = 0 \qquad \text{(iii)}$$

$$\sum_j n_{ij}\ \ \left(T_{oj} - k\mu - \sum_i n_{ij}\, \widehat{\alpha_i}\right) = 0 \qquad \text{(iv)}$$

From equation (iii):

$$T_{oo} = nk\mu + r \sum_{i=1}^{v} \alpha_i \qquad \text{(v)}$$

From equation (iv) we get:

$\sum_j n_{ij}T_{oj} = rk\mu + \sum_j n_{ij} \sum_h nj_j \widehat{\alpha_h}$

$$= rk\mu + \sum_j n_{ij}^2 \ \widehat{\alpha_\iota} + \sum_{h\neq i=1}^{v}\left(\sum_j n_{ij}\ n_{hj}\right)\widehat{\alpha_h}$$

$$= rk\mu + r\widehat{\alpha_\iota} + \sum_{h\neq i=1}^{v} \lambda\widehat{\alpha_h}$$

$or \sum_j n_{ij}\ T_{oj} = rk\hat{\mu} + r\alpha_i + \lambda(\sum_{h=i}^{v} \widehat{\alpha_h} - \alpha_i)$

Using the restrichtion $\sum_{i=1}^{v} \alpha_i = 0$, we have from above equation:

$\hat{\mu} = \frac{T_{oo}}{bk}$ $and\ T_i^{'} - rk\hat{\mu} = (r-\lambda)\hat{\alpha}_i\ where\ T_i = \sum_j n_{ij}\ T_{oj}$

$or\ \widehat{\alpha_\iota} = \frac{T_i' - rk\hat{\mu}}{(r-\lambda)} = \frac{T_i' - rk\frac{T_{oo}}{bk}}{r-\lambda} = \frac{T_i' - rk\frac{T_{oo}}{b}}{r-\lambda}$

Then $\hat{\alpha}_i$ is known as Inter-Block estimate.

ANOVA Table for Inter Block analysis of BIBD

Sources of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Blocks (adjusted)	b-1	$\sum_j \frac{{T_{oj}}^2}{k} + \sum_i Q_i\alpha_i - \sum_i \frac{{T_{oj}}^2}{r}$	$MSB = \frac{SSB}{b-1}$	$F_1 = \frac{MSB}{MSE}$
Treatment (unadjusted)	v-1	$\sum_{i=1}^{v} \frac{T_{io}^2}{r} - C.F.$	$MST = \frac{SST}{v-1}$	$F_2 = \frac{MST}{MSE}$
Error	bk-b-v+1	$\sum\sum h_{ij}\, y_{ij}^2 - \frac{1}{k}\sum_j T_{oj}^2 - \sum Q_i\alpha_i$	$MST = \frac{SSE}{bk-b-v+1}$	
Total	bk-1	$\sum h_{ij}y_{ij}^2 - CF$		

SSB (adj) + SST (un adj) = SSB (un adj) + SST (adj)

For finding an estimate of ${\sigma_b}^2$

$E[SSB(adj)] = E[SSB\ (adj)] = E\left[\sum_j \frac{{T_{oj}}^2}{k} + \frac{k}{\lambda v}\sum_i {Q_i}^2 - \sum_i \frac{{T_{oo}}^2}{r}\right],\ \left(where\ \widehat{\alpha_\iota} = \frac{k}{\lambda v} Q_i\right)$

Since terms on R.H.S do not involve μ and α_i hence for sampling we can assume that they are zero.

$$\therefore E(SSB) = \frac{1}{k}E\left(\sum_j T_{oj}{}^2\right) + \frac{k}{\lambda v}E\left(\sum_i Q_i{}^2\right) - \frac{1}{r}E\left(\sum_i T_{oo}{}^2\right)$$

$$= \frac{1}{k}\sum_j E\left(T_{oj}{}^2\right) + \frac{k}{\lambda v}\sum_i E\left(Q_i{}^2\right) - \frac{1}{r}\sum_i E\left(T_{oo}{}^2\right)$$

$$= \frac{1}{k}\sum_j V\left(T_{oj}\right) + \frac{k}{\lambda v}\sum_i v\,(Q_i) - \frac{1}{r}\sum_i v\,(T_{io})$$

$$= \frac{1}{k}\sum_j (k^2\sigma_b{}^2 + k\sigma_e{}^2) + \frac{k}{\lambda v}\sum_i r\frac{(k-1)}{k}\ \sigma_e{}^2 - \frac{1}{r}\sum_i (r\sigma_b{}^2 + r\,\sigma_e{}^2)$$

$$= \sum_j (\sigma_b{}^2 + \sigma_e{}^2) + r\frac{(k-1)}{k}\sum_i \sigma_e{}^2 = \sum_i (\sigma_b{}^2 + \sigma_e{}^2)$$

$V(T_{io}) = V\left(\sum_j n_{ij}\beta_j + \sum_i n_{ij}e_{ij}\right) = \sum_j n_{ij}\,v\left(\beta_j\right) + \sum_j n_{ij}\,v\left(e_{ij}\right)$

$$= r\sigma_b{}^2 + r\sigma_e{}^2$$

$E\left(T_{oj}\right) = E\left(k\beta_j + \sum n_{ij}\,e_{ij}\right) = 0$

(ii) we have $Q_i = 0 = E\left(Q_i\right) = 0$

(iii) we have $n_{ij}\,y_{ij} = n_{ij}\beta_j + n_{ij}c_{ij} = T_{io} = \sum_j n_{ij}\,y_{ij} = \sum_j n_{ij}\beta_j + \sum_j h_{ij}c_{ij}\ E(T_{io}) = 0$

$We\ assume\ \sum_i \alpha_i = \sum_j \beta_j = 0\ \ then\ v(T_{io}) = V(f_i)$

Now,

$$= bk\ \sigma_b{}^2 + b\sigma_e{}^2 + \frac{1}{\lambda v}\ vr(k+1)\sigma_e{}^2 - (v\sigma_b{}^2 + v\sigma_e{}^2)$$

$$= (bk - v)\sigma_e{}^2 + \left\{(b - v) + \frac{1}{\lambda}\ \lambda(v-1)\right\}\sigma_e{}^2$$

$$= (bk - v)\sigma_b{}^2 + (b-1)\sigma_e{}^2$$

$Therefore$

$$E\ \left(\frac{SSB(adj)}{b-1}\right) = E\left(MSB(adj)\right)$$

$$= \frac{bk-v}{b-1}\ \sigma_b{}^2 + \sigma_e{}^2$$

$$E(MSB - MSE) = \frac{bk-v}{b-1}\ \sigma_b{}^2 + \sigma_e{}^2 - \sigma_e{}^2 = \frac{bk-v}{b-1}\ \sigma_b{}^2$$

or

$$= E\left[\frac{b-1}{bk-v}\{MSB - MSE\}\right] = \sigma_b{}^2$$

$$Estimate\ of\ \sigma_b{}^2 = \widehat{\sigma_b}^2 = \frac{b-1}{bk-v}\{MSB - MSE\}$$

$$= \frac{b-1}{n-v}(MSB - MSE)$$

$where\ N = bk$

7.6 Resolvable BIBD

A BIBD with parameters v, r, b, k, and λ is said to be resolvable if the b blocks can be divided into r groups of b/r blocks each, b/r blocks forming any of these groups give a complete replication of all the v treatments.

(e.g., if b=8, r=4, b/r=2, i.e., 2 blocks occurring at least one)

Evidently in the case of resolvable design b is a multiple of r.

***Theorem*: -** In a resolvable BIBD with parameters v, r, b, k, and λ

$$\mathrm{b} \geq \mathrm{v} + \mathrm{r} - 1$$

***Proof:* -** Consider the incidence matrix N of this groups of columns each where any group of columns is such that; one occurs once and only once in each row of the group. By adding 1st, 2nd ------(b/r-1)th column to the b/rth column of a group we obtain a column consisting of one only, as there are groups and for each of these groups and column add it to the same vector. We have:

$v = rk\ of\ (\underline{N}) = r(\underline{N})\ and\ v \leq b - (r-1) = b - r + 1$

$$\#\ of\ v = c\ thin\ \underline{N} = \begin{pmatrix} 1\ 0\ 1\ 0\ 10 \\ 1\ 0\ 0\ 1\ 0\ 1 \\ 1\ 0\ 0\ 0\ 1\ 0 \\ 0\ 1\ 0\ 1\ 1\ 0 \\ 0\ 1\ 1\ 0\ 1\ 0 \\ 0\ 1\ 0\ 10\ 1 \end{pmatrix}$$

Another Point: - Since the design is resolvable b/r must be an integer, say equal to n. i.e., b/r = n, (say)

$\Rightarrow b = r.n$ (i)

But for a BIBD, vr = bk

$\Rightarrow vr = rnk \Rightarrow v = nk$ (ii)

Also, for a BIBD

$r\ (k-1) = \lambda(v-1) = \lambda\ (nk-1)$

$$\Rightarrow r = \frac{\lambda(nk-1)}{k-1} = \frac{\lambda\ nk-1}{k-1}$$

$$= \frac{\lambda nk-\lambda-\lambda n+\lambda n}{k-1} = \frac{\lambda n(k-1)+\lambda(n-1)}{k-1}$$

$$= \lambda n + \frac{\lambda(n-1)}{(k-1)}$$

$$\lambda n = \frac{\lambda(n-1)}{k-1} \qquad \text{(iii)}$$

Since r, λ and n are all integers, r- λn must be an integer and hence from equation (iii) must be an integer.

Now, if possible, let $b < v + r - 1$

$i.e. b - r < v - 1$ (iv)

$\Rightarrow rn - r = r(n-1) < v - 1;$ (from (i))

$$\Rightarrow r(n-1) < \frac{r(n-1)}{\lambda};\ [Since\ r(k-1) = \lambda(v-1)]$$

$$\Rightarrow n - 1 < \frac{k-1}{\lambda} \Rightarrow \frac{\lambda(n-1)}{k-1} < 1$$

Which is a contradiction to the fact that $\frac{\lambda(n-1)}{k-1}$ is natural number (+ve integer). Hence the assumption is wrong.

$$\therefore b \geq v + r - 1$$

7.7 Affine Resolvable BIBD

A resolvable design is said to be affine resolvable if b=r+v-1 and any two blocks from deferent sets have k^2/v treatments common where k^2/v is an integer.

Example: Consider the following data for the catalyst experiment and analyse it.

Treatment (Catalyst)	***Block (Batch of Raw Material)***				
	1	***2***	***3***	***4***	$y_{i.}$

1	73	74	__	71	218
2	__	75	67	72	214
3	73	75	68	__	216
4	75	__	72	75	222
$y_{.j}$	221	224	207	218	y.. = 870

Solution: This is a BIBD with a = 4, b = 4, k = 3, r = 3, λ = 2 and N = 12.

Total Sum of Square (TSS) = $\sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{12} = 63156 - \frac{870^2}{12} = 81$

Block Sum of Square (SSB) = $\frac{1}{3}\sum_{j=1}^{4} y_{.j}^2 - \frac{y_{..}^2}{12} =$

$$= \frac{1}{3}[221^2 + 207^2 + 224^2 + 218^2] - \frac{870^2}{12} = 55$$

To compute the treatment sum of squares adjusted for blocks, we first determine the adjusted treatment totals as:

$Q_1 = (218) - \frac{1}{3}(221 + 224 + 218) = \frac{-9}{3}$

$Q_2 = (214) - \frac{1}{3}(207 + 224 + 218) = \frac{-7}{3}$

$Q_3 = (216) - \frac{1}{3}(221 + 207 + 224) = \frac{-4}{3}$

$Q_4 = (222) - \frac{1}{3}(221 + 207 + 218) = \frac{20}{3}$

The adjusted sum of square for treatments is computed as:

SSTr (Adjusted) = $\frac{k\sum_{i=1}^{4} Q_i^2}{\lambda a} = \frac{3[(-9/3)^2+(-7/3)^2+(-4/3)^2+(20/3)^2]}{2*4} = 22.75$

Sum of Square due to Error (SSE) = TSS – SSTr (Adjusted) – SSB

= 81 – 22.75 – 55 = 3.25

The analysis of variance (ANOVA) is shown in table below:

Sources of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Treatments (adjusted for blocks)	*3*	*22.75*	*7.58*	*11.66*

Blocks	*3*	*55.00*	–	
Error	*5*	*3.25*	*0.65*	
Total	*11*	*81*		

Because the $F_{Tab.} = 5.4095 < F_{Cal.} = 11.66$, we conclude that the catalyst employed has a significant effect on the time of reaction.

7.8 Partially Balanced Incomplete Block Design (PBIBD)

Balanced incomplete block design which was studied earlier are the most efficient among all connected incomplete block design in which each block has the same which each block has treatment is number of plots and each treatment is replicated the same no. of time. However, BIBD do not always exists and for certain only with an extremely large no. of replicates.

Corrected Design

An incomplete block design in which all the treatment contracts are estimable is known as a connected design.

Partially balanced incomplete design (PBIBD) we used to overcome these difficulties. In this the number of replicates for each treatment can be made much smaller as compared. BIBD, However the design though connected is no longer balanced. i.e., All treatment contrasts of the type t_i-t_j are not estimate is the same variance.

Association Schemes

The concept of an association scheme is needed for the definition of PBIBD Given & symbols 1,2,------v. We have an association scheme with m classes is the following conditions ae satisfy.

(i) Any two symbols are either 1st, 2nd ----------- soon or mth association being symmetrical i.e. if the symbol α has n_i no. of ith associate, then β is the ith associate of α .

(ii) Each symbol α has n_i no. of ith associates the no. ni being independent of $\alpha(i = 1, ------, m)$

(iii) If α and β are the ith associates, the no. of symbols that are jth associates of c and kth associates of β, is p_{jk}^{i} and is independent of the pair of ith associates α and β .

Ex. Consider the following arrangement of six symbols i.e., i-----6.

1 2 3

4	5	6

With respect to each symbol the other symbols in the same row the first associates the one other symbol in the 2nd associate and the remaining two symbols are the IIIrd associate, 4th 2nd associate and 5, 6 are the 3rd associate of the 1(one).

Treatment	Ist associate	IInd associate	IIIrd associate
1	2,3	4	4,6
2	1,3	5	4,6
3	1,2	6	4,5
4	5,6	1	2,3
5	4,6	2	1,3
6	4,5	3	1,2

$p'_{23} = 1$ $\quad p'_{13} = 2$ $\quad \alpha = 1$ $\quad \beta = 4$

The result is a 3-class association scheme with $n_1 = 2$, $n_2 = 1$, $n_3 = 2$. The method can be used to generate 3 class associate schemes for m×n symbols by arranging in m rows and n columns such schemes are called rectangular association scheme.

Another example of a two-class association scheme is the triangular association scheme obtain by arranging $V = \frac{n(n-1)}{2}$ symbols in n rows and columns are as follows:

(i) The positions in the principal diagonal are left back.

(ii) N(n-1)/2 above the principal diagonal are filled by the numbers 1---- v corresponding to the symbols.

(iii) The position below the principal diagonal is filled so as to maintain symmetry above the principal diagonal.

The symbols entering in some row or columns with i are the first associate of i and rest are 2nd associate. Thus $V = 6 = \frac{n(n-1)}{2} \Longrightarrow n = 4$

×	1	2	3
1	×	4	5
2	4	×	6
3	5	6	×

The Ist and 2nd associates are as follows:

Treatment	Ist associate	IInd associate
1	2,3,4,5	6
2	1,3,4,6	5
3	1,2,2,6	4
4	1,5,2,6	3
5	1,3,4,6	2
6	2,3,4,5	1

PBIBD- Suppose the v treatments follow an m- association scheme, then we get a PBIBD with m association class if these v treatments are arrange into b blocks of size k(< V) such that:

(i) Every treatment occurs at most once in a block.

(ii) Every treatment occurs exactly r blocks

(iii) If two treatments α and β are the ith associates, then they occure together in λ_i blocks, the no being independent of the particular pair of ith associates α and β.

The numbers V, b,r, λ_i , n_i (i=1, ----,m) are known as the parameter of first kind and $p^i_{jk}(i,j,k = 1,----,m)$ are known as the parameter of 2nd kind.

7.8.1 Relationship among the Parameters of PBIBD

(i) vr = bk

Proof: **-** The R.H.S. given the total no. of plots in the design and the L.H.S. given the total no. of treatment which are to be used for the total no. of plots in the design consequently Vr = bk.

(ii) $\boldsymbol{\sum_{i=1}^{m} n_i = v - 1}$

Proof: - Let us consider any particular treatment α out of the v treatments and then we will be left this other v-1 treatments other than the one which was considered for m association scheme any treat will be either 1st associate or 2nd associate or so on up to nth associate of treat

α. Also, there are n_i no of i^{th} associate of treatment α, therefore v-1 will be equal to n_1+n_2+-----+nm.

(iii) $\boldsymbol{\sum_{i=1}^{m} n_i\lambda_i = r(k-1)}$

Where λ_i is the no. of times the i^{th} associate of treatment α occur together.

***Proof:* -** Let us consider r blocks in which a particular treat α occurs from this block we can found r (k-1) pairs of treatments α keeping as one of the treatments. Among these pairs, the i^{th} associate of α must occur λ_i times and there are n_i numbers of i^{th} associates of $\alpha(i = 1, - - - - nn)$. Hence $\sum_{i=1}^{m} n_i\lambda_i = r(k-1)$.

(iv) $\boldsymbol{\sum_{k=1}^{m} p_{jk}^{i} = n_j - \partial_{ij}; where\ \partial_{ij} = \begin{cases} 1; & if\ \ i = 1 \\ 0; & if\ \ i \neq j \end{cases}}$

***Proof:* -** Let α and β be the i^{th} associate. In this case, the k^{th} associates of $\alpha(k = 1 - - - - n)$ should cover all the n_j no. of j^{th} associate of $\beta. (j \neq i)$. Thus:

$\sum_{k=1}^{m} p_{jk}^{i} = n_j(x \neq j)\ i.e. p_{j1}^{i} + p_{j2}^{i} \pm - - - p_{jm}^{i} = n_j$

In the previous example: j=1, α =1, β = 2

$$p_{11}^{2} = 4, \qquad p_{12}^{2} = 0, \qquad n_1 = 4, \qquad i \neq j$$

When i=j, α itself will be one of the j^{th} associates of β. Hence the k^{th} associate of α (k=1----m), should cover all the n_j-1 associates of β. Thus

$\sum_{k=1}^{m} p_{jk}^{i} = n_j - 1\ if\ i = j$

Combining these we have the result: $\sum_{k=1}^{m} p_{jk}^{i} = n_j = \partial_{ij}$

(v) $\boldsymbol{n_i\, p_{jk}^{i} = n_j p_{ik}^{j}}$

***Proof:* -** Consider a treat α, let G_i denote the set of i^{th} associate of α. Each treatment in G_i has exactly p_{jk}^{i} kth associates in G_j.

$$\# \begin{bmatrix} \alpha = 1, \beta = 2 \quad G_i = \{2,3,4,5\}, \quad G_2 = \{b\} \\ j \quad G_i = \{1,3,4,6\}, \quad G_2 = \{5\} \\ k = 1,2 \\ p_{jk}^{i} \quad p_{21}^{1} = 0, \; p_{11}^{1} = 2 \end{bmatrix}$$

Similarly, each treatment in G_j has exactly p_{ik}^{j} k^{th} associates in G_i. Thus the no. of pairs of k^{th} associates that can be obtained by taking one treatment from G_i and another from G_j is n_i p_{jk}^{i} on the one hand and on the other hand $n_j p_{ik}^{j}$. Hence $n_i \, p_{jk}^{i} = n_j p_{ik}^{j}$.

(vi) The number of independent parameters of the 2nd kind p_{jk}^{i} $i, j, k = 1 - - - - m$ in the case of a PBIBD with m class (i.e., PBIBD(m)) is m(m2-1)/6

Proof: - Let

$$P_1 = \begin{pmatrix} p'_{11} & p'_{1m} \\ p'_{21} & p'_{2m} \\ \vdots & \vdots \\ \vdots & \vdots \\ p_{m1}^{1} & p'_{mm} \end{pmatrix}, - - - - P_i = \begin{pmatrix} p_{11}^{i} & p_{1m}^{i} \\ p_{21}^{i} & p_{2m}^{i} \\ \vdots & \vdots \\ \vdots & \vdots \\ p_{m1}^{i} & p_{mm}^{i} \end{pmatrix}, P_m = \begin{pmatrix} p_{11}^{m} & p_{1m}^{m} \\ p_{21}^{m} & p_{2m}^{m} \\ \vdots & \vdots \\ \vdots & \vdots \\ p_{m1}^{m} & p_{mm}^{m} \end{pmatrix}$$

We know that:

$$\sum_{k=1}^{m} p_{jk}^{i} = n_j = \partial_{ij} \qquad \text{(i)}$$

i.e., row tables of matrix Pi, i-1-----m are fixed (i.e., $n_j - \partial_{ij}$) Also we have

$$n_j p_{jk}^{i} = n_i p_{jk}^{i} \qquad \text{(ii)}$$

on taking i=1, j=2 we have $n_1 p_{2k}^{1} = n_2 p_{1k}^{2}$

i.e. Once being determined the elements of the 2^{nd} row of matrix P_1, then the elements of 1^{st} row of P_2 is known similarly if we know the 3^{rd} row of P_2, then the 2^{nd} row of P_3 is known i.e. for i=3, j=2 we have $n_2 p_{3k}^{2} = n_3 p_{2k}^{3}$ and soon further, we observe that matrix Pi, i=1,---m are symmetric matrixes (i.e. $p_{jk}^{i} = p_{jk}^{i}$), Let us consider the matrix P_1. Since P_1 is a symmetric matrix therefore the no. of independent parameters in the first row of P_1 is m, in the 2^{nd} row is (m-1), in the 3^{rd} row is (m-2) and so on and in the m^{th} row is 1(one). But we know from relation, we know that the row totals of Pi matrices are fixed.

Therefore, instead of m independent parameters in the I^{st} row of P_1, there will be only (m-1) independent parameters. Similarly in the 2^{nd} row of P_1 there will be (m-2) instead of (m-1) independent parameter and in the 3^{rd} row m=3 instead of m-2 independent parameter and so on and in the mth row zero independent parameters. Hence the no. of independent parameters in P_1 = (m-1) + (m-2) + ------+1+0 $= \frac{m(m-1)}{2} = m_{C_2}$.

Now consider the matrix P_2 and observing from (ii) we know that the elements of 1st row can be determine form these of elements of 2nd row of P_1. Hence there is no independent parameter in the first row of P_2. Agreeing P_1 we see that the no. of independent parameter in the 2nd row of P_2 in m-2, the no. of independent parameter in 3rd row P_2 is m-3 and no independent parameter in the mth row for P_2 = (m-2) + ------+1+0 $= \frac{(m-1)(m-2)}{2} = m - 1_{C_2}$.

Similarly considering P_3, we observe that the elements of the 1st row of P_3 can be determined from the elements of 3rd row of P, and the elements of 2nd row of P_3 from third row of P_2, which means that there is no independent parameter either in the first row or in the 2nd row of P_3. Arguing as about we see that the total no. of independent parameters in the 3rd row of P_3 is m-3, in the 4th row m-4 and so on and in the mth row no independent parameter in the row. Hence the no. of independent parameter in P_3 = (m-3) + ------+1+0 $= \frac{(m-2)(m-3)}{2} = m - 2_{C_2}$.

Arguing is the similarly manner the total no. of independent parameter in.

$$P_i = m - (i-1)_{C_2} = m - (i+1)_{C_2}$$

and

$$P_{m-1} = m - (m-2)_{C_2} = 2_{C_2} = 1, P_m = 0$$

i.e., there are no independent parameters in P_m. Hence the total no. of independent parameters in P_1, P_2, -----P_m is:

$$m_{C_2} + m - 1_{C_2} + m - 2_{C_2} + 3_{C_2} + 2_{C_2} = \sum_{i=0}^{m-2} m - i_{C_2} = \frac{m(m^2-1)}{6}$$

Proved.

7.9 Compounding BIBD

Let S_1, S_2, ………S_b be the block of a BIBD with parameters u, b, k, r and λ then if $S_i^{'}$ is said to be the complementary block of Si which contains elements i.e.

$$S_i^{'} = \{1, 2, ----, v\} - S_i$$

Then the design form with the blocks $S_1^{'}, S_2^{'} ----S_b^{'}$ will form another BIBD known as the complementary of the original design with parameters V_1 = V and b_1=b , but r_1, = b-r, k_1 = v-k and λ_1= b-2r + λ.

7.10 Complementary PBIBD

For every arithmetically possible design with parameter of the first kind (i.e. b, r, v, k, $\lambda, \lambda_2 --- \lambda_m, n_1, n_2, --- n_m$) and the associated parameters of the 2nd kind p^i_{jk} there is a complementary design with the same no. of blocks and of treatment as before but having (v-k) plots per block and b-r replications of each treatment the λ^s_i of the design will be (b-2r) more than the λ_i's of the first design.

The n_i's and the p^i_{jk} will be the same for both the designs.

7.11 Self-Assessment Exercise

1. Define a Balanced Incomplete Block Design (BIBD) and state the important relations among its parameters.
2. Describe the analysis of Balanced Incomplete Block Design (BIBD) without recovery of inter-block information.
3. Seven different hardwood concentrations are being studied to determine their effect on the strength of the paper produced. However, the pilot plant can only produce three runs each day. As days may differ, the analyst uses the balanced incomplete block design that follows. Analyze the data from this experiment (use α=0.05) and draw conclusions.

Hardwood Concentration (%)	*Days*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
2	114				120		117
4	126	120				119	
6		137	117				134
8	141		129	149			
10		145		150	143		
12			120		118	123	
14				136		130	127

7.12 Summary

This unit provides an overview of the concepts of Balanced Incomplete Block Design (BIBD) and Partially Balanced Incomplete Block Design (PBIBD), their construction and analysis as well as some different types of block designs such as resolvable design and affine resolvable design.

7.13 References

- Das, M.N. & Giri, N. (1979). *Design and Analysis of experiments*, Wiley Eastern.
- Gupta, V.K. & Nigam, A.K. (1978-79). *Handbook an analysis of Agriculture Experiment*, IASRI Publication.
- Montgomery, C.D. (1976). *Design and Analysis of Experiments*, Wiley, New York.
- Mukhopadhyay, P (2011). *Applied Statistics (2nd edition revised reprint*), Books & Allied (P) Ltd.

7.14 Further Reading

- Dean, A. & Voss, D. (1999). *Design and Analysis of Experiments*, Springer.
- Dey, A. (1986). *Theory of Block Designs*, Wiley Eastern.
- John P.W.M. (1971). *Statistical Design and Analysis of Experiments*, Mc Millan.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*, Wiley Eastern.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments*, Wiley, New York.
- Rao, C.R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, North Holland.
- Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*, Wiley.

UNIT-8 SPLIT & STRIP PLOT DESIGN

Structure

8.1 Introduction

The complete block designs, like completely randomised designs, randomized block designs and latin square designs are unsuitable if the number of varieties in varietal trials is large, say, exceeding ten. In factorial experiments, when the number of treatment combinations is large, the device of confounding is used to reduce the block size. The process ensures more precise estimation of lower order interactions at the cost of some of the less important higher order interactions which are confounded with the (incomplete) blocks. In varietal trials with no factorial structure of treatments comparison of all possible pairs of treatments are required to be estimated desirably with the same precision. Therefore, no contrast can be completely confounded with blocks as in factorial trials. It can be seen that precision of estimate of the difference between two treatments, more precisely effects of two treatments depends on the number of times these treatments occur in blocks, i.e., number of replications of two treatments. This fact has been used to construct designs for varietal or similar trials with large number of treatments so as to reduce the block size and hence obtain *incomplete block designs*. It is an experimental design, used to compare several treatments while dealing with constraints on resources, time, or space. This plan doesn't include every possible combination of treatments in every block and works well when it's not possible to run the whole set of treatments in each block, or when researchers want to lower the chance of mistakes and account for certain factors without completely randomizing the treatments. For example, consider a plant breeding experiment where different varieties of a crop need to be tested for yield. But because of room issues, it's not possible to grow all of them in the same place. With an incomplete block plan, varieties that need similar conditions for soil or climate could be put into blocks. Each block would have a subset of all the varieties. This design lets researchers test each variety while taking into account factors that are unique to each place.

Split-Plot Design and Strip-Plot Design are also two experimental design techniques used in scientific research, particularly in agriculture and industrial experiments. These designs are variants of the more common completely randomized design or randomized complete block design. A split-plot design is a type of experimental design that looks at the effects of more than one factor or treatment. It is especially useful when it's not possible to use all treatments on all experimental units because of time, space, or cost issues. This is often used in scientific study, like in agriculture, engineering, and industrial testing. Whereas, a strip-plot design is used to look into how different causes or treatments affect a population. This design is like a split-plot design, but it has some differences and can be used in different situations. For example, it can be used when it's not possible or cost-effective to apply all treatments to all experimental units and researchers want to control for certain factors or changes in the experimental area.

8.2 Objectives

After going through this unit, you should be able to:

- Perform the Inter-Block and Intra-Block analysis of Incomplete Block Designs,
- Construction and analysis of Split Plot Design,
- Understand the concept of Strip Plot Design.

8.3 Incomplete Block Design

Let y_{ij} denote the observation receiving i^{th} treatment in the j^{th} block and let n_{ij} be the number of such treatments. Then the observation table in case of an incomplete block design is given by:

Treatments	***Blocks***					
	1	***2***	$\cdots$	***j***	$\cdots$	***b***
1	$n_{11}y_{11}$	$n_{12}y_{12}$	$\cdots$	$n_{1j}y_{1j}$	$\cdots$	$n_{1b}y_{1b}$
2	$n_{21}y_{21}$	$n_{22}y_{22}$	$\cdots$	$n_{2j}y_{2j}$	$\cdots$	$n_{2b}y_{2b}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
i	$n_{i1}y_{i1}$	$n_{i2}y_{i2}$	$\cdots$	$n_{ij}y_{ij}$	$\cdots$	$n_{ib}y_{ib}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
v	$n_{v1}y_{v1}$	$n_{v2}y_{v2}$	$\cdots$	$n_{vj}y_{vj}$	$\cdots$	$n_{vb}y_{vb}$

Note that only k plots of each block is non-zero, while the rest of $(v - k)$ plots are zero. The mathematical model is then given by:

$$n_{ij}y_{ij} = n_{ij}\left(\mu + \alpha_i + \beta_j + e_{ij}\right),$$

where

μ is the general mean, α_i is the additive effect due to i^{th} treatment, β_j is the additive effect due to j^{th} block, $e_{ij}^{'}s$ are the random effects which are assumed to be *iid* random variables distributed according to $N(0, \sigma^2)$.

The mathematical model in case of an incomplete block design is a particular case of the model for a two-way classification with unequal number of observations per cell with no interaction between rows and columns and n_{ij} = 1 or 0. Here in the case of an incomplete block design, the additive effect β_j may be a fixed effect or a random effect having certain distribution. In case $\beta_j^{'}s$ are fixed effects, then we have ***Intra-Block Analysis*** (Analysis without recovery of inter-block information). If $\beta_j^{'}s$ are random effects having certain distribution, then we have ***Inter-Block Analysis*** (Analysis with recovery of inter-block information).

8.3.1 Intra-Block Analysis of an Incomplete Block Design

Intra block analysis is a special case of analysis of a two-way classification with unequal number of observations per cell. If n_{ij} denotes the number of observations in the $(i, j)^{th}$

cell of a two way classification, then the same analysis will hold for an incomplete block design with n_{ij} = 1 or 0 . The mathematical model is:

$$n_{ij}y_{ij} = n_{ij}\left(\mu + \alpha_i + \beta_j + e_{ij}\right),$$

where

μ is the general mean, α_i is the additive effect due to *i*th treatment, β_j is the additive effect due to *j*th block which is assumed to be fixed effect, $e_{ij}^{'}s$ are the random effects which are assumed to be *iid* random variables distributed according to $N(0,\sigma^2)$.

In a two-way classification with p rows and q columns and unequal number of observations per cell, we have:

SSRows (adjusted) = $\sum_{i=1}^{p} Q_i\alpha_i$,

where

$Q_i = T_{i..} - \sum_{j=1}^{q} \frac{n_{ij}T_{.j.}}{n_{.j}}$ and

$c_{i1}\alpha_1 + c_{i2}\alpha_2 + \cdots + c_{ip}\alpha_p = Q_i$ with $\alpha_1 + \alpha_2 + \cdots + \alpha_p = 0$.

Here we have *p* = *v*, *q* = *b*,

$n_{i1} + n_{i2} + \cdots + n_{ib} = \sum_{j=1}^{b} n_{ij} = r$

$n_{1j} + n_{2j} + \cdots + n_{vj} = \sum_{i=1}^{v} n_{ij} = k$.

Also, since $n_{ij}^2 = n_{ij}$ = 1 or 0,

$c_{ii} = n_{i.} - \sum_{j=1}^{b} \frac{n_{ij}^2}{n_{.j}} = n_{i.} - \sum_{j=1}^{b} \frac{n_{ij}}{k} = n_{i.} - \frac{n_{i.}}{k} = r - \frac{r}{k} = r\left(1 - \frac{1}{k}\right) = \frac{r(k-1)}{k}$ and

$c_{il}(i \neq l) = -\sum_{j=1}^{b} \frac{n_{ij}n_{lj}}{n_{.j}} = -\frac{1}{k}\sum_{j=1}^{b} n_{ij}n_{lj} = -\frac{1}{k}\lambda_{il}$, where λ_{il} is the number of times the *i*th and the *l*th treatments occur together in all the blocks.

Thus, we have,

$$\frac{r(k-1)}{k}\alpha_i - \sum_{l(\neq i)=1}^{v} \frac{\lambda_{il}}{k}\alpha_l = Q_i \text{ and } \alpha_1 + \alpha_2 + \cdots + \alpha_v = 0. \qquad \text{(A)}$$

Now $T_{i.} = \sum_{j=1}^{b} n_{ij}y_{ij}$ is the total yield for the *i*th treatment,

$T_{.j} = \sum_{i=1}^{v} n_{ij}y_{ij}$ is the total yield for the *j*th block and

$Q_i = T_{i.} - \sum_{j=1}^{q} \frac{n_{ij}T_{.j}}{n_{.j}}$ is the adjusted yield for the *i*th treatment.

The adjustment being that we subtract the block average for those blocks where in the i^{th} treatment occurs from the total yield of the i^{th} treatment. Hence the Sum of Squares due to Treatments (adjusted) is equal to $\sum_{i=1}^{v} Q_i\alpha_i$ and α_i is to be determined from the set of equations given in (A).

ANOVA Table for Intra block analysis of an Incomplete Block Design

Sources of Variation	*Degrees of freedom*	*Sum of Squares*	*Mean Sum of Squares*	*Variance Ratio*
Treatment (adjusted)	$v-1$	$S.S.T. = \sum_{i=1}^{v} Q_i\alpha_i$	$M.S.T. = \frac{S.S.T.}{v-1}$	$F = \frac{M.S.T}{M.S.E}$
Blocks (unadjusted)	$b-1$	$S.S.B. = \frac{1}{k}\sum_{j=1}^{b} T_{.j}^2 - C.F.$	$M.S.B. = \frac{S.S.B.}{b-1}$	
Error	$bk-v-b+1$	$S.S.E. = \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij}^2 - \frac{1}{k}\sum_{j=1}^{b} T_{.j}^2 - \sum_{i=1}^{v} Q_i\alpha_i$	$M.S.E. = \frac{S.S.E}{bk-v-b+1}$	
Total	$bk-1$	$T.S.S. = \sum_{i=1}^{v}\sum_{j=1}^{b} n_{ij}y_{ij}^2 - C.F.$		

8.3.2 Inter-Block Analysis of an Incomplete Block Design

The mathematical model is then given by:

$$n_{ij}y_{ij} = n_{ij}\left(\mu + \alpha_i + \beta_j + e_{ij}\right) \tag{1}$$

where μ is the general mean, α_i is the additive effect due to i^{th} treatment, $\beta_j^{'}s$ are the additive effect due to j^{th} block which are assumed to be *iid* random variables distributed according to $N(0,\sigma_b^2)$, $e_{ij}^{'}s$ are the random effects which are assumed to be *iid* random variables distributed according to $N(0,\sigma_e^2)$. Also $\beta_j^{'}s$ are independent if $e_{ij}^{'}s$.

Summing (1) over i, we get:

$\sum_{i=1}^{v} n_{ij}y_{ij} = \sum_{i=1}^{v} n_{ij}\left(\mu + \alpha_i + \beta_j + e_{ij}\right)$.

Or $T_{.j} = \sum_{i=1}^{v} n_{ij}\mu + \sum_{i=1}^{v} n_{ij}\,\alpha_i + \sum_{i=1}^{v} n_{ij}\,\beta_j + \sum_{i=1}^{v} n_{ij}\,e_{ij}$ [since $\sum_{i=1}^{v} n_{ij} = k$]

$= k\mu + \sum_{i=1}^{v} n_{ij}\,\alpha_i + k\beta_j + \sum_{i=1}^{v} n_{ij}\,e_{ij}$

$= k\mu + \sum_{i=1}^{v} n_{ij}\,\alpha_i + f_j,$

where $f_j = k\beta_j + \sum_{i=1}^{v} n_{ij}\, e_{ij}$ are residuals, which are assumed to be *iid* variables each distributed as $N(0, \sigma_f^2)$, and

$$\sigma_f^2 = V(f_j) = V(k\beta_j + \sum_{i=1}^{v} n_{ij}\, e_{ij})$$

$$= k^2 V(\beta_j) + \sum_{i=1}^{v} n_{ij}^2 V(e_{ij})$$

$$= k^2\sigma_b^2 + \sum_{i=1}^{v} n_{ij}\sigma_e^2 = k^2\sigma_b^2 + k\sigma_e^2.$$

For estimating μ and α_i, we differentiate the sum of squares due to residuals

$$\sum_{j=1}^{b} f_j^2 = \sum_{j=1}^{b} \left(T_{.j} - k\mu - \sum_{i=1}^{v} n_{ij}\, \alpha_i\right)^2$$

with respect to μ and α_i and equate to zero individually.

Thus,

$$\frac{\delta\left(\sum_{j=1}^{b} f_j^2\right)}{\delta\mu} = -2k\sum_{j=1}^{b}\left(T_{.j} - k\mu - \sum_{i=1}^{v} n_{ij}\, \alpha_i\right) = 0.$$

Or $\sum_{j=1}^{b}\left(T_{.j} - k\mu - \sum_{i=1}^{v} n_{ij}\, \alpha_i\right) = 0$

$\Rightarrow\ T_{..} - bk\mu - \sum_{i=1}^{v}\left(\sum_{j=1}^{b} n_{ij}\right)\alpha_i = 0.$

Or $T_{..} = bk\mu + r\sum_{i=1}^{v} \alpha_i$ [since $\sum_{j=1}^{b} n_{ij} = r$].

Now since $\sum_{i=1}^{v} \alpha_i = 0$, we get an estimate of μ as $\hat{\mu} = \frac{T_{..}}{bk} = \bar{y}_{..}$.

Similarly, $\frac{\delta\left(\sum_{j=1}^{b} f_j^2\right)}{\delta\alpha_i} = 0$ gives:

$$\sum_{j=1}^{b} n_{ij}\left(T_{.j} - k\mu - \sum_{i=1}^{v} n_{ij}\, \alpha_i\right) = 0.$$

Or $\sum_{j=1}^{b} n_{ij} T_{.j} = k\sum_{j=1}^{b} n_{ij}\mu + \sum_{j=1}^{b} n_{ij}\sum_{h=1}^{v} n_{hj}\alpha_h$

$$= k\sum_{j=1}^{b} n_{ij}\mu + \sum_{j=1}^{b} n_{ij}^2\, \alpha_i + \sum_{h(\neq i)=1}^{v}\left(\sum_{j=1}^{b} n_{ij} n_{hj}\right)\alpha_h$$

$$= k\sum_{j=1}^{b} n_{ij}\mu + \sum_{j=1}^{b} n_{ij}\, \alpha_i + \sum_{h(\neq i)=1}^{v} \lambda_{ih}\, \alpha_h$$

$$= k\sum_{j=1}^{b} n_{ij}\bar{y}_{..} + r\alpha_i + \sum_{h(\neq i)=1}^{v} \lambda_{ih}\, \alpha_h.$$

This equation has to be solved for α_i for obtaining an estimate of α_i, which is simpler if all the $\lambda_{ih}^{'}s$ are same. That will be the case of a Balanced Incomplete block design.

8.4 Spilt Plot Design

In field experiment, sometimes a factor has to be applied to a large experimental unit. This is true when different types of ploughing or irrigation have to be compared. In such cases, it is possible to introduce a second factor which does not require large plots, with a small number of levels into the same experiment, at a little extra cost. This is done by splitting the plots (called whole plots) of the first factor into as many subplots as there are levels of the second factor.

A split-plot design with an RBD for the first set of treatments (called whole-plot treatments) is done by applying the whole-plot treatments to the plots of each block and then randomizing. The second set of treatments (called the sub-plot treatments) is then applied to each of the whole-plots of the blocks.

The difference between the split-plot design and an ordinary two-factor experiment is that, while in the former case the randomization is done separately for the whole-plot treatments (of a block) and the sub-plot treatments (of a whole-plot), while in the latter all the treatment combinations of the factors are allotted at random to the plots of a block.

This enables us to test for the main effects of sub-plot treatments and the interaction of the whole-plot treatment and the sub-plot treatments more efficiently than the main effects of the whole-plot treatments in a split-plot design. On the other hand, the main effects and the interaction are all tested equally efficiently in the two-factor experiment in an RBD.

There is another interpretation of the split-plot design which brings out its similarity with a confounded design. If the sub-plots are considered as plots and the whole plots as blocks, we find that the differences among the whole plots are same as the differences among the levels of whole-plot treatments. So, this design may be said to have confounded the main effects of the whole-plot treatments. In this respect this design violates the earlier recommendation in that confounding in factorial experiments should preferably restricted to higher-order interactions.

8.4.1 Layout

The p levels of the factor A are randomized according to the plan used in an RBD or an LSD. The q levels of the factor B are then randomized inside each whole-plot of factor A by dividing each whole plot into q sub plots. The randomization is carried out separately for each whole-plot of a block.

8.4.2 Analysis

Suppose we have a factor A at p levels, which are arranged in an RBD using r blocks, and a second factor B at q levels are applied to each plot of a block after subdividing each plot into q subplots. So, there are p whole-plots in a block, q subplots in a whole-plot and there are r replications.

The mathematical model used is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ij} + \gamma_k + \delta_{jk} + e_{ijk}; \qquad 1)$$

$i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, p;\ k = 1, 2, \ldots, q,$

where μ is the general mean, α_i is the additive effect due to i^{th} replication, β_j is the additive effect due to j^{th} whole-plot, γ_k is the additive effect due to k^{th} sub-plot of j^{th} whole-plot, δ_{jk} is the additive effect due to interaction between to j^{th} whole-plot and k^{th} sub-plot, e_{ij} is the error for the RBD with only whole plot treatments and e_{ijk} is the error for the entire design, which are assumed to be *iid* variates each distributed as N $(0, \sigma^2)$. The side conditions are:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_{\substack{j \\ \forall k}} \delta_{jk} = \sum_{\substack{k \\ \forall j}} \delta_{jk} = \sum_{\substack{i \\ \forall j}} e_{ij} = \sum_{\substack{j \\ \forall i}} e_{ij} = 0.$$

Then unrestricted residual sum of squares is given by:

$$\text{SSE} = \sum_i \sum_j \sum_k e_{ijk}^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right)^2.$$

To estimate the parameters, we differentiate this with respect to $\mu, \alpha_i, \beta_j, e_{ij}, \gamma_k$ and δ_{jk} and equate to zero individually. Thus, differentiation with respect to μ gives:

$$\sum_i \sum_j \sum_k \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0$$

$$\Rightarrow T_{\ldots} - pqr\mu = 0 \ \Rightarrow \ \hat{\mu} = \frac{T_{\ldots}}{pqr} = \bar{y}_{\ldots}.$$

Differentiating with respect to α_i gives:

$$\sum_j \sum_k \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0, \ \forall\, i = 1, 2, \ldots, r.$$

$$\Rightarrow T_{i..} - pq\mu - pq\alpha_i = 0 \Rightarrow \hat{\alpha}_i = \frac{T_{i..}}{pq} - \hat{\mu} = \bar{y}_{i..} - \bar{y}_{\ldots}.$$

Differentiating with respect to β_j gives:

$$\sum_i \sum_k \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0, \ \forall\, j = 1, 2, \ldots, p.$$

$$\Rightarrow T_{.j.} - qr\mu - qr\beta_j = 0 \Rightarrow \hat{\beta}_j = \frac{T_{.j.}}{qr} - \hat{\mu} = \bar{y}_{.j.} - \bar{y}_{\ldots}.$$

Differentiating with respect to e_{ij} gives:

$\Sigma_k \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0, \ \forall\, i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, p$.

$\Rightarrow T_{ij.} - q\mu - q\alpha_i - q\beta_j - qe_{ij} = 0 \Rightarrow \hat{e}_{ij} = \frac{T_{ij.}}{q} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$

$= \bar{y}_{ij.} - \bar{y}_{...} - \bar{y}_{i..} + \bar{y}_{...} - \bar{y}_{.j.} + \bar{y}_{...} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$.

Differentiating with respect to γ_k gives:

$\Sigma_i \Sigma_j \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0, \ \forall\, k = 1, 2, \ldots, q.$

$\Rightarrow T_{..k} - pr\mu - pr\,\gamma_k = 0 \Rightarrow \hat{\gamma}_k = \frac{T_{..k}}{pr} - \hat{\mu} = \bar{y}_{..k} - \bar{y}_{...}$.

Differentiating with respect to δ_{jk} gives:

$\Sigma_i \left(y_{ijk} - \mu - \alpha_i - \beta_j - e_{ij} - \gamma_k - \delta_{jk}\right) = 0.$

$\Rightarrow T_{.jk} - r\mu - r\beta_j - r\gamma_k - r\delta_{jk} = 0 \Rightarrow \hat{\delta}_{jk} = \frac{T_{.jk}}{r} - \hat{\mu} - \hat{\beta}_j - \hat{\gamma}_k$

$= \bar{y}_{.jk} - \bar{y}_{...} - \bar{y}_{.j.} + \bar{y}_{...} - \bar{y}_{..k} + \bar{y}_{...} = \bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}.$

Hence, on substituting these estimates in (1), we get:

$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + \left(\bar{y}_{.j.} - \bar{y}_{...}\right) + \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right) + (\bar{y}_{..k} - \bar{y}_{...})$

$+ \left(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}\right) + e_{ijk}.$

$\Rightarrow \hat{e}_{ijk} = y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.}.$

Let us call the sum of squares of this error (which is due to chance cause) as SSE_2. Then:

$SSE_2 = \Sigma_i \Sigma_j \Sigma_k e_{ijk}^2 = \Sigma_i \Sigma_j \Sigma_k \left(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.}\right)^2.$

This will have $pqr - pr - pq + p = p(q-1)(r-1)$ degrees of freedom.

Now consider:

$y_{ijk} - \bar{y}_{...} = \left(\bar{y}_{ij.} - \bar{y}_{...}\right) + \left(y_{ijk} - \bar{y}_{ij.}\right).$

Squaring both sides and summing over i, j and k, we have

$\Sigma_i \Sigma_j \Sigma_k \left(y_{ijk} - \bar{y}_{...}\right)^2 = \Sigma_i \Sigma_j \Sigma_k \left(\bar{y}_{ij.} - \bar{y}_{...}\right)^2 + \Sigma_i \Sigma_j \Sigma_k \left(y_{ijk} - \bar{y}_{ij.}\right)^2$

Or

Grand Total Sum of Squares (GTSS) = TSS (between whole plots) + TSS (between sub-plots within whole plots)

Also $\bar{y}_{ij.} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})$

Squaring both sides and summing over *i*, *j* and *k*, we get:

$$\sum_i \sum_j \sum_k \left(\bar{y}_{ij.} - \bar{y}_{...}\right)^2 = \sum_i \sum_j \sum_k (\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_i \sum_j \sum_k \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2 + \sum_i \sum_j \sum_k \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right)^2$$

TSS (between whole plots) = S.S.Replicate + S.S. Whole plot + SSE_1,

where $SSE_1 = \sum_i \sum_j \sum_k \hat{e}_{ij}^2$, represents the sum of squares due to error (or chance cause) in case of an ordinary RBD without splitting the plots.

This will have *rp – 1 = (r – 1) + (p – 1) + (rp – r – p + 1)* degrees of freedom.

Again:

$$y_{ijk} - \bar{y}_{ij.} = (\bar{y}_{..k} - \bar{y}_{...}) + \left(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}\right) + \left(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.}\right).$$

Squaring both sides and summing over *i*, *j* and *k*, we get

$$\sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{ij.}\right)^2 = \sum_i \sum_j \sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + \sum_i \sum_j \sum_k \left(\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}\right)^2 +$$

$$\sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.}\right)^2$$

TSS (between sub-plots within whole plots) = S.S. Sub-plot + S.S. Interaction (Whole-plot × Sub-plot) + SSE_2.

This will have $rp(q-1) = (q-1) + (pq-p-q+1) + (pqr-rp-pq+p)$

$= (q-1) + (p-1)(q-1) + p(q-1)(r-1)$ degrees of freedom.

Hence the analysis of variance table is given by:

8.4.3 ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***
Replication *Whole-plot* E_1	$r-1$ $p-1$ $(r-1)(p-1)$	SSR $= pq\sum_{i=1}^{r}(\bar{y}_{i..} - \bar{y}_{...})^2$ SSWp $= qr\sum_{j=1}^{p}(\bar{y}_{.j.} - \bar{y}_{...})^2$ SS E_1 $=\sum_i \sum_j \sum_k (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	MSR= $\frac{SSR}{r-1}$ MSWp $=\frac{SSWp}{p-1}$ MS E_1 = $\frac{SSE_1}{(r-1)(p-1)}$	 $F_1 = \frac{MSWp}{MS\ E1}$
Total between whole-plots	$rp-1$	TSS (between whole plots) = $\sum_i \sum_j \sum_k (\bar{y}_{ij.} - \bar{y}_{...})^2$		
Sub-plots *Interaction (Whole-plot × Sub-plot)* E_2	$q-1$ $(p-1)(q-1)$	S.S. Sp = $rp\sum_{k=1}^{q}(\bar{y}_{..k} - \bar{y}_{...})^2$ SSI = $r\sum_j \sum_k (\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...})^2$ SSE_2 =	M.S. Sp = $\frac{S.S.Sp}{q-1}$ MSI = $\frac{SSI}{(p-1)(q-1)}$ MSE_2 = $\frac{SSE_2}{p(q-1)(r-1)}$	$F_2 = \frac{M.S.Sp}{MSE_2}$ $F_2 = \frac{MSI}{MSE_2}$

	$p(q-1)(r-1)$	$\sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{.jk} + \bar{y}_{.j.}\right)^2$		
Total between sub-plots within whole-plots	$pr(q-1)$	TSS (between sub-plots within whole plots) $= \sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{ij.}\right)^2$		
Grand total	$pqr-1$	GTSS $= \sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{...}\right)^2$		

Example: Consider a paper manufacturer who is interested in three different pulp preparation methods (the methods differ in the amount of hardwood in the pulp mixture) and four different cooking temperatures for the pulp and who wishes to study the effect of these two factors on the tensile strength of the paper. Each replicate of a factorial experiment requires 12 observations, and the experimenter has decided to run three replicates. This will require a total of 36 runs. The experimenter decides to conduct the experiment as follows. A batch of pulp is produced by one of the three methods under study. Then this batch is divided into four samples, and each sample is cooked at one of the four temperatures. Then a second batch of pulp is made up using another of the three methods. This second batch is also divided into four samples that are tested at the four temperatures. The process is then repeated, until all three replicates (36 runs) of the experiment are obtained. The data are shown in table below:

Pulp Preparation Method	*Replicate-1*			*Replicate-2*			*Replicate-3*		
Temperature (°F)	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
200	30	34	29	28	31	31	31	35	32
225	35	41	26	32	36	30	37	40	34
250	37	38	33	40	42	32	41	39	39
275	36	42	36	41	40	40	40	44	45

Solution: In this split-plot design we have 9 whole plots, and the preparation methods are called the whole plot or main treatments. Each whole plot is divided into four parts called subplots (or split-plots), and one temperature is assigned to each. Temperature is called the subplot treatment.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + e_{ijk}$$

$i = 1, 2, \ldots, r; j = 1, 2, \ldots, p; k = 1, 2, \ldots, q$

where α_i, β_j and $(\alpha\beta)_{ij}$ represent the whole plot and correspond, respectively, to replicates, main treatments (factor A), and **whole-plot error** (replicates x A), and γ_k, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, and $(\alpha\beta\gamma)_{ijk}$ represent the subplot and correspond, respectively, to the subplot treatment (factor B), the replicates x B and AB interactions, and the **subplot error** (replicates x AB). Note that the whole-plot error is the replicates x A interaction and the subplot error is the three-factor interaction replicates x AB. The sums of squares for these factors are computed as in the three-way analysis of variance without replication.

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***Variance Ratio***	
				$F_{Cal.}$	$F_{Tab.}$
Replicates	2	77.55	38.78		
Preparation method (A)	2	128.39	64.2	7.08	6.94
Whole plot error (replicates A)	4	36.28	9.07		
Temperature (B)	3	434.08	144.69	41.94	3.49
Replicates B	6	20.67	3.45		
AB	6	75.17	12.53	2.96	3.00
Subplot error (replicates AB)	12	50.83	4.24		
Total	35	822.97			

8.5 Strip Plot Design

In split plot design some factor requires smaller unit as compare to other and we can increase the precision on the factor B and factor AB such sacrifices some precision of A. in that design the factor A is lesser importance. However, sometimes we may have factors A and B each requiring larger units

***Example*:** If we compare different agricultural experiments and different spacing. Therefore, to accommodate both the factors in large units we use strip-plot design

In this design we divide each replicate (Block) into a number of rows (same as number of levels of one factor i.e. A) and a number of columns (same as number of levels of other factors i.e., B). The rows and columns are called strips

Let p levels of A are randomized in p rows and q levels of b in q columns of a replicate. Here a single entire column receives a single level of B. The allocation of A and B to the rows and columns will be a fresh for each of their replicates. Here since both factors are applied to the strips i.e., larger plots so the main effect of A and B will have lower precision as compared to interaction AB.

In this design there are three errors for different effects.

The ANOVA will be based on the following model:

$$y_{(ijk)} = \mu + \pi_i + \alpha_j + (\pi\alpha)_{(ij)} + \beta_k + (\pi\beta)_{(ik)} + (\alpha\beta)_{(jk)} + (\pi\alpha\beta)_{(ijk)}$$

Where,

$y_{(ijk)}$ be the yield of plot receiving $j^{(th)}$ level of A and $k^{(th)}$ level of b in the $i^{(th)}$ replicates

$(\pi\alpha)_{(ij)}$, $(\pi\beta)_{(ik)}$, $(\pi\alpha\beta)_{(ijk)}$ are error which are independently normally distributed with mean zero and variance σ_1^2 , σ_2^2 and σ_3^2 respectively.

α_j, β_k, $(\alpha\beta)_{(jk)}$ are fixed effect.

8.5.1 ANOVA Table

Sources of Variation	***Degrees of freedom***	***Sum of Squares***	***Mean Sum of Squares***	***E(MS)***
Replication	$r - 1$	SSR	MSR	
Treatment	$p - 1$	SSA	MSA	$\sigma_{e''}^2 + q\sigma_e^2 + (rq/p-1)\sum \tau_j^2$
Error I $(R \times A)$	$(r-1)(p-1)$	SSE_1	MSE_1	$\sigma_{e''}^2 + q\sigma_e^2$
Treatment (B)	$(q-1)$	SSB	MSB	$\sigma_{e''}^2 + p\sigma_{e'}^2 + (rp/q-1)\sum \gamma_k^2$
Error II	$(r-1)(q-1)$	SSE_2	MSE_2	$\sigma_{e''}^2 + p\sigma_{e'}^2$

$(R \times B)$				
Interaction $(A \times B)$	(p−1)(q−1)	SS(AB)	$MS(AB)$	$\sigma_{e''}^2 + r/(p-1)(q-1) + \Sigma \sum \delta_{(jk)}^2$
Error $II(R \times A \times B)$	$(r-1)(p-1)(q-1)$	SSE_3	MSE_3	$\sigma_{e''}^2$
Total	$(rpq-1)$			

8.6 Self-Assessment Exercise

1. Define Incomplete Block Design (IBD).
2. What is the need of Split Plot in a block design?
3. What do you mean by Split Plot Design? How will you do the analysis of Split Plot Design?
4. A process engineer is testing the yield of a product manufactured on three machines. Each machine can be operated at two power settings. Furthermore, a machine has three stations on which the product is formed. An experiment is conducted in which each machine is tested at both power settings, and three observations on yield are taken from each station. The runs are made in random order, and the results are shown in below table. Analyze this experiment, assuming that all three factors are fixed.

Station	*Machine-1*			*Machine-2*			*Machine-3*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
Power Setting 1	34.1	33.7	36.2	31.1	33.1	32.8	32.9	33.8	33.6
	30.3	34.9	36.8	33.5	34.7	35.1	33.0	33.4	32.8
	31.6	35.0	37.1	34.0	33.9	34.3	33.1	32.8	31.7
Power Setting 2	24.3	28.1	25.7	24.1	24.1	26.0	24.2	23.2	24.7
	26.3	29.3	26.1	25.0	25.1	27.1	26.1	27.4	22.0
	27.1	28.6	24.9	26.3	27.9	23.9	25.3	28.0	24.8

8.7 Summary

This unit provides an overview of the Inter-Block and Intra-Block analysis of Incomplete Block Designs (IBD), detailed analysis of Split Plot Designs and a brief introduction to Strip Plot Design.

8.8 References

- Das, M.N. & Giri, N. (1979). *Design and Analysis of experiments*, Wiley Eastern.

- Gupta, V.K. & Nigam, A.K. (1978-79). *Handbook an analysis of Agriculture Experiment*, IASRI Publication.
- Montgomery, C.D. (1976). *Design and Analysis of Experiments,* Wiley, New York.
- Mukhopadhyay, P (2011). *Applied Statistics (2nd edition revised reprint*), Books & Allied (P) Ltd.

8.9 Further Reading

- Dean, A. & Voss, D. (1999). *Design and Analysis of Experiments*, Springer.
- Dey, A. (1986). *Theory of Block Designs*, Wiley Eastern.
- John P.W.M. (1971). *Statistical Design and Analysis of Experiments*, Mc Millan.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*, Wiley Eastern.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments*, Wiley, New York.
- Rao, C.R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, North Holland.
- Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*, Wiley.

UNIT:9 OTHER ADVANCED DESIGN

Structure

9.1 Introduction

Advanced experimental designs such as such as Dual Block Design, Lattice Design, Cross-Over Design, Response Surface Design, Mixture Experiments etc. are sophisticated and specialized techniques used in scientific research to efficiently investigate the effects of multiple factors or treatments while controlling for various sources of variation.

In dual block designs, two different randomised full block designs are used together in one experiment which helps when there are two important factors, and each has its own set of control factors. Whereas, a linked block design is employed when the experimental units are organized into blocks with a specific structure or pattern. In clinical studies, cross-over designs are often used to see how different treatments affect the same group of people, often over more than one time period. Each person in the study gets each medicine at a different time.

Lattice designs are a way to make some types of resolvable incomplete block designs. These can be Balanced Incomplete Block (BIB) designs or Partially Balanced Incomplete Block (PBIB) designs, but some are not. Incomplete block designs are good and helpful when there are a lot of treatments and/or full blocks are not available or are not the right shape. Lattice designs were first used in large-scale agricultural studies (Yates, 1936) to compare a lot of different types of plants. Lattice designs are particularly valuable when resources are limited.

Response surface experiments are usually used in latter stages of experimentations, or after the important factors have been found, i.e., when a small group (usually between two and eight) of continuous factors that have been found as active are involved. It's used to show how the relationship between the factors and the outcome is curved. It lets us determine how to set our factors so that they have the least or most impact on a response or so that they optimize a specific objective. The design needs at least three levels for the factors in order to predict the curve. Because of this, reaction surface designs can get very big if the number of factors isn't limited. The goal is to make processes work better by making a model that can predict how the factors will affect the answer.

Mixture experiments are a type of response surface experiment. The goal of these experiments is to find more optimal response by combining ingredients in a certain way, either to maximize or minimize some property. For example, in construction of stainless steel, which is made up of Fe, Cu, Cr, and Ni, the qualities of the steel depend on the amounts of each component. The value of a factor is how much of the blend it makes up. Its value is between 0 and 1. There are at least three factors in a mixture experiment, and the sum of their amounts is one (100%). Because of this, its experimental space is usually triangle and the shape of a simplex.

9.2 Objectives

After going through this unit, you should be able to:

- Understand the basic concepts of Dual and Linked block designs, lattice design and cross-over design,
- Get an overview of optimal criteria for an optimal design, robust parameter design,
- Know the concept of Response Surface Design, Weighing Design and Mixture Experiments.

9.3 Dual Block Design

A design obtained by interchanging the columns of the design by row, if the design is D, then dual would be $D^{'}$

Let d be a connected GD design

- A connected design is one which allows estimability of every treatment contrast. if you can reach from one treatment to other through links or chain of alternate links and track.

$\vartheta \rightarrow treatments, b \rightarrow blocks, r_1, r_2 \ldots . r_v (replicates)$ $k_1, k_2 \ldots \ldots k_b$(Block size) and incidence matrix N.

$$R = diag(r_1, r_2 \ldots \ldots r_v)$$

$$k = diag(k_1, k_2 \ldots . k_b)$$

$$\mathrm{c} = \mathrm{R} - \mathrm{NK}^{(-1)}\mathrm{N}'$$

A design is connected iff

Rank $(c) = \vartheta - 1$

$c\underline{z} = \underline{Q}$

CI = C*(1 1 ….1)’ = 0

$Rank(c) < \vartheta - 1$

Let d be a connected G.D. design and let N_d be the incident matrix of d, then the eigen values of $N_d N_d^{'}$ and their multiplicities are:

$\theta_0 = rk, \alpha_0 = 1$

$\theta_1 = r - \lambda_1, \alpha_1 = m(m-1)$

$\theta_n = rk - \vartheta\lambda_2, \alpha_1 = m(m-1)$

$\alpha_i^{'}$s are multiplicities and are integers (here we observe that $N_d N_d^{'}$ is real symmetric and hence $N_d N_d^{'}$ is positive definite and hence θ_1’s should be positive)

For the GD designs, these can be classified into 3 groups

i. Singular if $r = \lambda_1$

ii. Semi-Singular if $r > \lambda_1 and\ rk - \vartheta\lambda_2 = 0$

iii. Regular if $r > \lambda_1$ and $rk > \vartheta\lambda_2$

9.4 Linked Block Design

An incomplete block design is called a linked block design (LBD) if every pair of blocks intersects in a constant number of treatments.

A square matrix A of order $n * n$ is called completely symmetric if $A = (\alpha - \beta)I_n + bJ_n$ for some scalars α *and* β .I_n is an identity matrix of order n and j_n is an $n * n$ Matrix of all ones.

$$\begin{pmatrix} \alpha & \beta & \beta & \cdots\cdots & \beta \\ \beta & \alpha & \beta & \cdots\cdots & \beta \\ . & . & . & \cdots\cdots & . \\ . & . & . & & . \\ \beta & \beta & \beta & \cdots\cdots & \alpha \end{pmatrix}$$

If $A_{(n*n)}$ is complete symmetric, then the eigne values of A are $(\alpha - \beta)$ with multiplicity $(n-1)$ and $\alpha + (n-1) * \beta$ with multiplicity 1.

$$A = (\alpha - \beta)I_n + \beta J_n$$

$$AI_n = \alpha + (\mathrm{n} - \mathrm{p})\beta I_n$$

Hence $det(A) = (\alpha - \beta)^{(n-1)}\alpha + (n-1)\beta$

If A is complete symmetric and invertible then $A^{(-1)}$ is also complete symmetric and is given by $A^{(-1)} = (c-d)I_n + dJ_n$

$where \quad c = \frac{(\alpha+(n-2)*\beta)}{(\alpha-\beta)\{(\alpha+(n-1)\beta))}$ and $d = \frac{(\beta)}{(\alpha-\beta)\{(\alpha+(n-1)\beta))}$

Also $AA^{(-1)} = I_n$

Lemma: A symmetric matrix A of order n is c.s. iff, A has only two distinct eigen values, one of these with multiplicity (n-1) and I_n is an eigenvector corresponding to other eigen value.

Proof: Let A be a symmetric matrix of order n and let A have only two distinct eigen values θ_1 *and* θ_2. Where θ_2 has multiplicity n-1 and 1_n is an eigen vector corresponding to θ_1, then their exist an orthogonal matrix

$$u = \left(n^{-\frac{1}{2}} \;\; 1_n'\right) such\ that$$

$$uAu' = \begin{pmatrix} \theta_1 & 0' \\ 0 & \theta_2 I_{n-1} \end{pmatrix}$$

$\therefore A = u'uAu'u = \theta_1 n^{-1} J_n + \theta_2 P'P$

$$= (\theta_1 - \theta_2) n^{-1} J_n + \theta_2 I_n \qquad Since\ \{P'P = I_n - n^{-1} J_n\}$$

The converse is also true because of the previous result.

For any matrix A the non-zero eigen values of AA' and A'A are the same including the multiplicities.

Result: Let d be a PBIBD with $m(\geq 2)$ associates classes and parameters $v, b, r, k, \lambda_i, n_i, p^i_{jj}\ (i, j, s = 1,2, \dots \dots \dots m)$. Then d is linked block design iff $N_d N_d'$ has only two non-zero eigen values $\theta_0 = rk$ and θ_1, with respective multiplicities $\alpha_0 = 1\ and\ \alpha_1 = b - 1$ (N_d is incidence matrix of d).

Proof: Let d be a linked block design with intersection number u,

then N'd*Nd =

$$\begin{pmatrix} k & \mu & \mu & \dots\dots & \mu \\ \mu & k & \mu & \dots\dots & \mu \\ . & . & . & \dots\dots & . \\ . & . & . & & . \\ \mu & \mu & \mu & \dots\dots & k \end{pmatrix}$$

$$(k - u) I_n + u J_n$$

Therefore, the eigen values of $N_d' N_d$ are $\theta_1 = k - u$ with multiplicity b-1 and $\theta_0 = K + (b - 1)u$ with multiplicity 1.

k = (b-1)*u is obtained by taking row sum.

Now

$N_d' N_d\ 1_d\ = k + (b - 1)u$

$\therefore \{k + (b - 1)u\} 1_n = N_d' N_d 1_v$

$$= r N_d' 1_v$$

$$= rk\ 1_n$$

$k + (b - 1)u = rk$

Conversely; let d be a PBIBD with block size K and replication r such that $N_d N_d'$ has only two non-zero eigen values $\theta_0 = rK$ with multiplicity 1 and $\theta_1$with multiplicity b-1.

Then $N_d N_d^{'}$ has only two non-zero eigen values θ_0 *and* θ_1, the multiplicity of $\theta_1$being b-1 (since $N_d^{'} N_d$ is a $b \times b$ matrix). Also, since $N_d^{'} N_d^{'} 1_b = rK\, 1_b; 1_b$ *is an eigenvector* corresponding to θ_0.

Thus, improving the lemma, we see that $N_d^{'} N_d$ is c.s. and is given by:

$$N_d^{'} N_d = (K - \theta_1) I_b + \theta_1 I_b$$

This shows that all the non-diagonal elements of N_d are θ_1 hence 'd' is a linked block design.

9.5 Lattice Design

Consider a BIBD with k^2 (i.e., $v=k^2$) treatments (k is block size) arranging b = k (k+1) blocks with k runs per block and r= k+1 replicates. Such a design is called a balanced Lattice. An example is shown in the following table for $k^2 = 9$ treatments in 12 blocks of three runs each.

Notice that the blocks can be grouped into sets such that each set contains a complete replicate. The analysis of variance for the balance Lattice design proceeds like that go BIBD, except that the S.S. for replicates is computed and removed from the S.S, for blocks replicate will have k d.f. and blocks will have k^2-1 d.f.

Table: - A 3 × 3 balanced Lattice design.

Block	***Replicate-I***	***Block***	***Replicate-II***
1	1 2 3	4	1 4 7
2	4 5 8	5	2 5 8
3	7 8 9	6	3 8 9
Block	***Replicate-III***	***Block***	***Replicate-IV***
7	1 5 9	10	1 8 6
8	7 2 6	11	4 2 9
9	4 8 3	12	7 5 3

Lattice Designs are frequently used in situations where there is a large number of treatment combinations. In order to reduce the size of the design, the experimenter may resort to partially balanced Lattices we briefly described some of these designs here. Two replicates of a design for k^2 treatments in 2k blocks, (i.e., $v=k^2$, b= 2k) of k runs each is called a simple Lattice, e.g., consider the first two replicates of the above design. The partial balance is easily seen, as for example treatment 2 appears in the same block with treatment 1, 3, 5, & 8, but does not appear at all with treatment 4, 6, 7 and 9. A Lattice design with k^2 treatment in 3k blocks (i.e., b = 3k) grouped into three replicates is called a triple Lattice. An example would be the I^{st} three replicates in the above table.

A Lattice arranged in 4 replicates is called a quadruple Lattice.

There are a number of other types of Lattice design that occasionally true useful for example the cubic Lattice design can be used for v=k3 treatments in k^2 blocks (i.e., b = k^2) of k runs each. A Lattice design for v=k (k+1) treatments in (k+1) blocks (i.e., b= (k+1)) of size k is called a rectangular Lattice.

9.6 Cross-Over Design

In some cases, it is desirable to apply two or more treatments consecutively to the same experimental unit. In dairy animal husbandry, for example, two or more feeds maybe given one after the other to the same animal. Experiments which involve the sequential application of several treatments to the same experimental unit are called change-over-trials.

Suppose we want to compare the effect of two treatments a and b, over time periods 1 and 2, viz, the intake of a at 1 followed by that of b at 2 (i.e., the sequence a-b) with the intake of b at 1 followed by that of a at 2 (the sequence b-a), on a number of experimental units. When two or more treatments are applied in sequence to the same experimental unit it is reasonable to assume that the response to the second and subsequent treatments maybe affected by the preceding treatments in the sequence. Such conditioning effects are called carry-over or residual effects. The carry-over effect can be minimized or eliminated by allowing a long enough rest period between administration of two treatments. Thus, carry-over effects may or may not exist depending on the nature of experiment. When they exist, we assume that the carry-over effects in any period are entirely due to the immediately preceding treatment. Thus, if the treatments are applied in the sequence a-b-c, then at period 2 there is a carry-over effect r_a of a and at period 3 there is a carry-over effect r_b of b only. Our problem is to estimate and test the direct $(\tau_a, \tau_b, \ldots.)$ and carry-over effects (r_a, r_b, …). The sum of direct effect and carry-over effect is called permanent effect or equilibrium effect of the treatment.

Consider the simplest case of two treatments a and b used in a trial over two periods. Each unit receives the treatments either in the sequence a-b or b-a. We divide the n experimental units randomly into two groups of sizes n_1, n_2 (n_1+n_2=n) and apply the sequence a-b to the members of the first group and b-a to the remaining units.

The periods should be of equal length. Note that since the same unit is given both the treatments the individual differences are eliminated in this design.

For $n = 6$, $n_1 = n_2 = 3$, the design may be as follows:

Period	***Units***					
	1	*2*	*3*	*4*	*5*	*6*
1	a	b	b	a	b	a
2	b	a	a	b	a	b

Ordinarily, the same number of units is assigned to each group since it minimizes the mean square due to error.

Let y_{ijk} = value of y on the j[th] unit in the i[th] sequence group at period k (k=1,2,.., j=1,2,…n_i, i=1,2). The data are shown in the following table:

Period	*Treatment*	*Data in Group-1*				*Sum*
1	a	y_{111}	y_{121}	……..	$y_{(1n_1 1)}$	G_{11}
2	b	y_{112}	y_{122}	……..	$y_{(1n_1 2)}$	G_{12}
Period	*Treatment*	*Data in Group-2*				*Sum*
1	b	y_{111}	y_{121}	……..	$y_{(2n_2 1)}$	G_{11}
2	a	y_{212}	y_{222}	……..	$y_{(2n_2 2)}$	G_{22}

Let $G_1 = G_{11} – G_{12} = G_{21} – G_{22}$

$D_{1j} = y_{1j1} – y_{1j2}$, j = 1,2,… n_1

$D_{2j} = y_{2j1} – y_{2j2}$, j = 1,2,… n_2

Sum of Square due to Treatments (SSTr) = $\frac{(n_2G_1 - n_1G_2)^2}{2nn_1n_2}$, $n = n_1+n_2$

Sum of Square due to Error (SSE) = $\frac{1}{2}\sum_i \sum_j D_{ij}^2 - \frac{1}{2n_1}G_1^2 - \frac{1}{2n_2}G_2^2$

The treatment means are: $\bar{y}_a = \bar{\bar{y}} + d \text{ and } \bar{y}_b = \bar{\bar{y}} - d$

Where,

$\bar{\bar{y}} = \frac{\sum_i \sum_k G_{ik}}{2n}$, $d = \frac{n_2G_1 - n_1G_2}{4n_1n_2}$

$V(\bar{y}_a) = V(\bar{y}_b) = \frac{n*\sigma^2}{4n_1n_2}$ $V(\bar{y}_a - \bar{y}_b) = \frac{n*\sigma^2}{2n_1n_2}$

A modification of cross-over design is a *switch-back or double-reversal design*. Here, each experimental unit receives either the treatment sequence a-b-a or the treatment sequence b-a-b over three periods 1,2,3. Periods should be of equal length. The n units are divided at random into two groups n_1, n_2 and the members of the first group receive the sequence a-b-a and the remaining the other sequence.

Let as before y_{ijk} denote the value of y on the j^{th} unit in the i^{th} sequence group at period k (i=1,2; j=1,2,…n_i, k=1,2,3).

Let

$G_{ik} = \sum_{j=1}^{n_i} y_{ijk}$;

$G_i = G_{i1} - 2G_{i2} + G_{i3}$, i=1,2; k=1,2,3

$D_{ij} = y_{ij1} - 2y_{ij2} + y_{ij3}$

Here,

Sum of Square due to Treatments (SSTr) = $\frac{(n_2G_1 - n_1G_2)^2}{6nn_1n_2}$

Sum of Square due to Error (SSE) = $\frac{1}{6}\sum_i \sum_j D_{ij}^2 - \frac{1}{6n_1}G_1^2 - \frac{1}{6n_2}G_2^2$

The treatment means are:

$\bar{\bar{y}} = \frac{1}{2n}\left[\sum_{i=1}^{2}\sum_{j=1}^{3} G_{ik} - \frac{(n_1 - n_2)(n_2G_1 - n_1G_2)}{8n_1n_2}\right]$

$\bar{y}_a = \bar{\bar{y}} + d \; and \; \bar{y}_b = \bar{\bar{y}} - d$

Where,

$d = \frac{n_2G_1 - n_1G_2}{8n_1n_2}$, $\quad V(\bar{y}_a) = V(\bar{y}_b) = \frac{3n*\sigma^2}{16n_1n_2}$, $\quad V(\bar{y}_a - \bar{y}_b) = \frac{3n*\sigma^2}{8n_1n_2}$

9.7 Optimal Designs for Response Surfaces

A lot of people use the basic response surface designs because they are pretty general and adaptable. A standard response surface design will usually work for a case where the experimental area is either a cube or a sphere. However, there are times when an experimenter may not think that a normal response surface design is the best option. In these situations, you could also think about optimal structures. There are several situations where some type of computer-generated design may be appropriate, such as:

a. ***Irregular Experimental Region*** - As long as the area of interest isn't a cube or a sphere, normal designs might not be the best choice. It's pretty common for areas of interest to be not straight. For instance, a researcher is looking into how a certain adhesive works. The glue is put on two pieces, and then it cures at a high temperature. The amount of glue used and the temperature at which it cures are the two things that matter. On a normal coded variable scale, these two factors run from -1 to +1. The experimenter knows that if too little glue is used and the cure temperature is too low, the parts will not stick together properly. In terms of the coded variables, this leads to a *constraint* on the design variables, say: $-1.5 \leq x_1 + x_2$, where x_1 represents the application amount of adhesive and x_2 represents the temperature. Furthermore, if the temperature is too high and too much adhesive is applied, the parts will be either damaged by heat stress or an inadequate bond will result. Thus, there is another constraint on the factor levels: $x_1 + x_2 \leq 1$.

b. ***Nonstandard Model*** **-** Experimenters usually choose a first- or second-order response surface model, knowing that this model is only close to the real process at work. But sometimes the person doing the experiment may know something unique about the process being studied that makes them think of a model that isn't the usual one. For example, the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{112} x_1^2 x_2 + \beta_{1112} x_1^3 x_2 + \varepsilon$$

c. ***Unusual Sample Size Requirements*** - In some cases, an experimenter may need to cut down on the number of runs that a standard response surface plan calls for. Let's say we want to fit a second-order model with four factors. Depending on how many center points are chosen, the central composite plan in this case needs between 28 and 30 runs. The model, on the other hand, only has 15 words. The experimenter will want a plan with fewer trials if the runs cost a lot of money or take a long time. Computer-made images can be used for this, but there are other ways to do it as well. For instance, a small composite design can be made for four factors with 20 runs and four center points, and a hybrid design can be made with as few as 16 runs. These options might be better than using a computer-made design to cut down on the number of tries.

9.7.1 Design Optimality Criteria

1. D-Optimal Criterion

It is the most widely used optimality criterion, and a design is said to be D-optimal if: $\left|\left(X'X\right)^{-1}\right|$ is minimized. A D-optimal design minimizes the volume of the joint confidence region on the vector of regression coefficients. A measure of the relative efficiency of design 1 to design 2 according to the D-criterion is given by $D_e = \left(\frac{\left|\left(X_2'X_2\right)^{-1}\right|}{\left|\left(X_1'X_1\right)^{-1}\right|}\right)^{1/p}$, where X_1 and X_2 are the X matrices for the two designs and p is the number of model parameters.

2. A-Optimality Criterion

It deals with only the variances of the regression coefficients. A design is A-optimal if it minimizes the sum of the main diagonal elements of $(X'X)^{-1}$. (This is called the trace of $(X'X)^{-1}$, usually denoted $tr((X'X)^{-1})$. Thus, an A-optimal design minimizes the sum of the variances of the regression coefficients.

3. G-Optimality Criterion

Because many response surface experiments are concerned with the prediction of the response, *prediction variance criteria* are of considerable practical interest. The most popular of these is the G-optimality criterion. A design is said to be G-optimal if it minimizes the maximum scaled prediction variance over the design region; that is, if the maximum value of $\frac{NV[\hat{y}(x)]}{\sigma^2}$ over the design region is a minimum, where N is the number of points in the design. If the model has p parameters, the G-efficiency of a design is just $G_e = \frac{p}{max\frac{NV[\hat{y}(x)]}{\sigma^2}}$.

4. V-Optimality Criterion

The V-criterion considers the prediction variance at a set of points of interest in the design region, say $x_1, x_2, \ldots, x_m$. The set of points could be the candidate set from which the design was selected, or it could be some other collection of points that have specific meaning to the experimenter. A design that minimizes the average prediction variance over this set of m points is a V-optimal design.

The D criteria are usually thought to be the best for first-order designs because they deal with parameter estimates, which is very important for screening, which is where the first-order model is most often used. Since the G and I criteria are prediction-based, they would most likely be used for second-order models. This is because second-order models are often used for optimisation, and good prediction qualities are necessary for optimisation. The I criteria is easier to use than the G criteria, and it can be found in a number of software programmes.

9.8 Response Surface Design

Suppose that we are interested in improving the yield of a chemical process we know that from a result of a characterization experiment the two most important process variables that influence the yield are operating temperature and reaction time. Suppose a process currently runs at 145°F and 2.1 hours of reaction time producing the yield about 80%.

Contour Plot

A contour plot is a series of lines or curves that identify values of the factors for which the response is constant. Curves for several (usually equal spaced) values of the response are plotted. These contours are projection on the time-temperature region of cross section of the yield surface. This surface is called "*Response Surface*".

"A Response surface is a geometrical representation obtained when a response variable is plotted as a function of one or more quantitative data (factors)."

The response surface is unknown to process personnel, so experimental method will be required to optimize the yield with respect to time and temperature. For this, we perform a experiment that varies in time and temperature together i.e., a factorial experiment. One we found the region of optimum, second experimental would typically be performed. The objective this experiment is to develop an empirical model of the process and to obtain a more precise estimate of the optimum operating condition for time and temperature. This approach to process optimization is called Response Surface Methodology.

9.8.1 Response Surface Methodology

The Response Surface Methodology (RSM) is a collection of mathematical and statistical techniques useful for developing, improving and optimizing processes.

OR

The RSM is a collection of mathematical and statistical techniques that are useful for the modeling and analysis of problems in which a response of interest is influenced by several variable and the objective is to optimize this response.

The RSM was developed by Box and Willson (1951). It also has important application in design development and formulating new product as well as in the improvement of existing product design. The most extensive applications of RSM are in the industrial world, particularly, in situations when several input variables potentially influence some performance measure or quality characteristic of the product or process. It is typically measured on a continuous scale. Another application of RSM; it involves more than one response. The input variables are called Independent Variables.

In most of RSM problems, the relationship between response variable and independent variable is unknown. So, the first step in RSM, is to find a suitable approximation for the true relationship between response variable and independent variables. In many cases either a first order as second order model is used.

A fixed order model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdots\cdots + \beta_k x_k + \epsilon$$

Where,

y = Response variable

$x_i's$ independent variable,

ϵ is a random error distributed as normally with mean 0 and variance σ^2,

$\beta_0, \beta_1, \dots\dots.. \beta_k$ are the parameters.

E(y) = $\eta = \beta_0 + \sum_i \beta_i x_i$

And if there is a curvature (i.e., interaction relationship) in the system, then we use the second order model which as

$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i \le j \le k} \beta_{ij} x_i x_j + \epsilon$

$= \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i \le j \le k} \beta_{ij} x_i x_j + \epsilon$

Almost all RSM problems use one or both of these models.

To find the estimate of parameters β_i and β_{ij} method of least square is used. The response surface analysis is then performed using fitted surface. If the fitted surface is an adequate approximation of the true response function; then the analysis of yield surface will be approximately equivalent to analysis of actual system. The model parameters can be estimated most effectively if proper experimental designs are used to collect the data. The design for fitted response surface are called Response Surface Design.

Some Features of a desirable Designs

1. It provides a responsible distribution of data points (and hence information) throughout the region of interest.
2. Allows model adequacy, including lack of fit.
3. Allows experiments to be performed in blocks.
4. Allows designs of higher order to be built up sequentially.
5. Provides on interval estimate of error.
6. Provides precise estimates of the model coefficients.
7. Provides a good profile of the prediction variance throughout the experimental region.
8. Provides responsible robustness against outliers or missing value.
9. Does not require a large number of runs.
10. Does not require too many levels of the independent variables.
11. Ensure the simplicity of calculation of model parameters.

9.8.2 Fitting of First Order Model

Let $X_1, X_2, \dots\dots\dots, X_\nu$ denotes the ν independent factors. Let x_{iu} denotes the i^{th} factor level in u^{th} combination (i=1,2,…….,ν) and $\mathrm{u} = 1,2,\dots, \mathrm{N}$, where $\mathrm{N} = $ total treatment combinations) and let y_u denotes the response obtained from the u^{th} combination of factors. The first order model is

$$y_u = \beta_0 + \sum_{i=1}^{\nu} \beta_i x_{iu} + \epsilon_u \quad ; u = 1,2,\dots\dots N \tag{1}$$

Where ϵ_u is error and distributed or normal with mean 0 and variance σ^2, i.e., $\epsilon_u \sim i.i.d. N(0, \sigma^2)$.

$\beta_i^{'}s$ are the parameter and they are to be estimated by method of least squares. Let $b_0, b_i^{'}s$ denotes the estimates of parameter, then

Error sum of squares = $\sum_u (y_u - b_0 - \sum_{i=1}^{\nu} b_i x_i)^2$ (2)

Differentiating equation (2) with respect to $b_0, b_1, b_2, \dots\dots\dots\dots b_\nu$ partially and equating to zero, we get the following normal equations.

$$\left.\begin{array}{l}
\sum_{u=1}^{N} y_u = N\, b_0 + b_1 \sum_{u=1}^{N} x_{1u} + b_2 \sum_{u=1}^{N} x_{2u} + \cdots + b_\nu \sum_{u=1}^{N} x_{2u} \\
\sum_{u=1}^{N} x_{1u} y_u = b_0 \sum_u x_{1u} + b_1 \sum_u x_{1u}^2 + b_2 \sum_u x_{1u} x_{2u} + \cdots + b_\nu \sum_u x_{1u} x_{\nu u} \\
. \\
. \\
.
\end{array}\right\} \tag{3}$$

$\sum_{u=1}^{N} x_{iu}\, y_u \;=\; b_0 \sum_u x_{iu} \;+ b_1 \;\sum_u x_{1u} x_{iu} \;+ \cdots + b_v \;\sum_u x_{iu} x_{vu}$

.

.

.

$\sum_u x_{vu} y_u = b_0 \sum_u x_{vu} x_{1u} + b_1 \sum_u x_{vu} x_{1u} + \cdots b_v \sum_u x_{vu}^2$

Now use of simplification condition

$\sum_{u=1}^{N} x_{iu} \;= 0 \quad \forall\, i = \;1,2, \dots \dots\,,\; v$

$\sum_{u=1}^{N} x_{iu}^2 = N \quad \forall\, i = \;1,2, \dots \dots ., v$

$\sum_{u=1}^{N} x_{iu} x_{ju} \;= 0 \quad \forall\, i \neq j\,; j\,, i = \;1,2, \dots \dots, v$

Now from (3) and using simplification conditions we get,

$\sum_u y_u = N b_0$

$b_0 = \frac{1}{N} \sum_{u=1}^{N} y_u$

And,

$\sum_u x_{iu} y_u = b_i \sum_u x_{iu}^2$

$b_i = \frac{\sum_u x_{iu} y_u}{\sum_u x_{iu}^2} \;=\; \frac{\sum_{u=1}^{2} x_{iu} y_u}{N}$

Now the variance of these estimators are:

$Var(b_0) \;= Var\,(\frac{1}{N} \sum_u y_u)$

$= \frac{1}{N^2} \;\sum_u Var(y_u)$

$= \frac{1}{N^2} \;\sum_u \sigma^2$

$Var(b_0) = \frac{\sigma^2}{N}$ = Coefficient of $\sum_u y_u$ in the expression of $b_0 \sigma^2$

And

$Var\;(b_i) = Var\;(\frac{1}{N} \sum_u x_{iu} y_u)$

$= \frac{1}{N^2} \;\sum_u x_{iu}^2 Var(y_u)$

$= \frac{1}{N^2} . N\,. \sigma^2$

$Var(b_i) = \frac{\sigma^2}{N}$ = Coefficient of $\sum_u y_u x_{iu}$ in the expression of $b_i \sigma^2$

Note:

1. Covariance terms are zero as error terms are uncorrelated.
2. Plackett Busman (1946) Design can be used to fit this model along with its simplifying conditions.

9.8.3 Rotatability

As the number of factors increases the 3^k factorial become inefficient and impractical. These experiments need a large number of observations. For example: 3^5=243 and 3^{10}=59049. Further these designs do not give equal precession of fitted response at points (factor level combination). Further these designs do not give equal precision for fitted response at points (factor level combination) that are at equal distance from the center of factor space. A design that has this property is termed as a rotatable design.

When fitting a specified response surface model, A design is rotatable if fitted model estimates the response with equal precision at all points in the factor space that are equidistance from the center of the design.

All 2^k completely factorials are rotatable but 3^k factorials are not.

9.8.4 Blocking

When using the response surface designs, it is often necessary to consider blocking to eliminate nuisance variables. Such problem may occur when a higher order, say second-order design is assembled sequentially from lower order, say. Such necessity arises due to various reasons. For example, considerable time may elapse between the running of the first-order design and the running of the supplemental experiments which are required to build up a second-order design, and during this time, the test conditions may change which makes necessary to use blocking.

9.8.5 Orthogonality

A response surface design is said to be block orthogonally if it is divided into blocks such that block effects do not affect the parameter estimates of the response surface model. If a 2^k

or 2^{k-p} design is used as a first-order response surface design, the center points in these designs should be allocated among the blocks.

For a second-order design to block orthogonally, two conditions must be satisfied. If there are n_b observations in the b^{th} block, then these conditions are:

1. Each block must be a first-order orthogonal design; that is,

$$\sum_{u=1}^{n_b} x_{iu}x_{ju} = 0;\ i \neq j = 0,1,\ldots k\ for\ all\ b$$

Where x_{iu} and x_{ju} are the levels of i^{th} and j^{th} variable in the u^{th} run of the experiment with $x_{0u} = 1$ for all u.

2. The fraction of the total sum of squares for each variable contributed by every block must be equal to the fraction of the total observations that occur in the block, that is:

$$\frac{\sum_{u=1}^{n_b} x_{iu}^2}{\sum_{u=1}^{N} x_{iu}^2} = \frac{n_b}{N};\ i = 1,2,\ldots.,for\ all\ b$$

Where, N is the number of runs in the design.

9.9 Weighing Design

A weighing design can be formally defined as, given p objects to be weighted in groups in N weights a weighing design consists of N grouping of the p objects such that in each grouping the p objects are made into three sets of sizes $n_1, n_2\ and\ n_0$ and while weighing, the set of size n_1 is placed on one pan, say the left one that of size n_2 on the other pan and the third of size n_0 is admitted from the weighing .there will thus be one weighing for each of the N groupings.

Suppose, a chemical engineer wants to find out the weights of 5 objects:

Objects	O_1	O_2	O_3	O_4	O_5
True Object	w_1	w_2	w_3	w_4	w_5

We want to weigh the chemical balance (2 pans). Suppose each record weight has variance σ^2. Thus, we use the standard technique to weight.

Suppose we take the 6 weighing and weight of object = weight of object including weight of pan -weight of pan

Let Y_i denotes the reading in the $i^{(th)}$ object and y_0 reading without any objects. Then estimate weight of the $i^{(th)}$ object

$$\widehat{w_i} = y_i - y_0$$

$V(\widehat{w_i}) = V(y_i - y_0) = 2\sigma^2$

And Standard Error $(S.E.) = \sqrt{2}.\sigma^2$

Now a question arise, can we reduce this standard error?

Yates (1935), observe that when measurement are made on the sets of objects rather than weighing them individually. The accuracy of the measurement of weights increases. In the above example, we can weigh them in sets of n objects rather than weighing them individually with 5 objects we have $\binom{5}{4}$,such set viz.

w_1	w_2	w_3	w_4
w_1	w_2	w_3	w_5
w_1	w_2	w_4	w_5
w_1	w_3	w_4	w_5
w_2	w_3	w_4	w_5

w_0(Weight of the pan)$= y_0$ (1)

With the above setting we have the following equation

$w_0 + w_1 + w_2 + w_3 + w_4 = y_1$ (2)

$w_0 + w_1 + w_2 + w_3 + w_5 = y_2$ (3)

$w_0 + w_1 + w_2 + w_4 + w_5 = y_3$ (4)

$w_0 + w_1 + w_3 + w_4 + w_5 = y_4$ (5)

$w_0 + w_2 + w_3 + w_4 + w_5 = y_5$ (6)

We solve this equation to get the estimate of true weights:

For w_i, add equation 2) 3) 4) and 5), we get

$4w_0 + 4w_1 + 3(w_2 + w_3 + w_4 + w_5) = y_1 + y_2 + y_3 + y_4$

$4w_0 + 4w_1 + 3(w_0 + w_2 + w_3 + w_4 + w_5) - 3w_0 = y_1 + y_2 + y_3 + y_4$

$w_0 + 4w_1 + 3y_5 = y_1 + y_2 + y_3 + y_5$

$\hat{w} = \left((y_1 + y_2 + y_3 + y_4) - 3y_5 - y_0\right)/4$

For w_2, add equations 2), 3), 4), and 6), we get

$4w_0 + 4w_2 + 3(w_1 + w_3 + w_4 + w_5) = y_1 + y_2 + y_3 + y_5$

$4w_0 + 4w_1 + 3(w_2 + w_3 + w_4 + w_5) - 3w_0 = y_1 + y_2 + y_3 + y_5$

$w_0 + 4w_2 + 3y_4 = y_1 + y_2 + y_3 + y_5$

$\widehat{w_2} = \left((y_1 + y_2 + y_3 + y_5) - 3y_4 - y_0\right)/4$

Similarly, we get:

$\widehat{w_3} = \left((y_1 + y_2 + y_4 + y_5) - 3y_3 - y_0\right)/4$

$\widehat{w_4} = \left((y_1 + y_2 + y_3 + y_5) - 3y_2 - y_0\right)/4$

& $\widehat{w_5} = \left((y_1 + y_2 + y_3 + y_4) - 3y_1 - y_0\right)/4$

& $var(w_i) = (1 + 1 + 1 + 9 + 1)\sigma^2/16 = 14\sigma^2/16 = 7\sigma^2/8$

Hoteling (1994) suggested that one can improve the weighing of the Yates techniques:
Equation will become

$w_0 - w_1 - w_2 - w_3 - w_4 - w_5 = y_0$ (1)

$w_0 + w_1 + w_2 + w_3 + w_4 - w_5 = y_1$ (2)

$w_0 + w_1 + w_2 + w_3 - w_4 + w_5 = y_2$ (3)

$w_0 + w_1 + w_2 - w_3 + w_4 + w_5 = y_3$ (4)

$w_0 + w_1 - w_2 + w_3 + w_4 + w_5 = y_4$ (5)

$w_0 - w_1 + w_2 + w_3 + w_4 + w_5 = y_5$ (6)

To get the estimates of w_1,we add all those equation in which w occurs with positive sign we get

$4w_0 + 4w_1 + 3(w_2 + w_3 + w_4 + w_5) - (w_2 + w_3 + w_4 + w_5) = y_1 + y_2 + y_3 + y_4$

$4w_0 + 4w_1 + 3(w_0 - w_1 + w_2 + w_3 + w_4 + w_5) - 3w_0 + 3w_1 + (w_0 - w_1 - w_2 - w_3 - w_4 - w_5) - w_0 + w_1) = y_1 + y_2 + y_3 + y_4$

$8w_1 + 3y_5 + y_0 = y_1 + y_2 + y_3 + y_4$

$\widehat{w_1} = (y_1 + y_2 + y_3 + y_4 - y_0 - 3y_5)/8$

Similarly, we can obtain other w_i,then

$v(\widehat{w_i}) = (1 + 1 + 1 + 1 + 1 + 9)\sigma^2/164 = 14/64\sigma^2$

And S.E $.(\widehat{w_1}) = \sigma\sqrt{7}/32 < \sigma * \sqrt{7}/8$

9.10 Mixture Experiments

In this experiment, the response depends only on the proportion of the ingredients present the mixture and is not a function of the amount of the mixture.

Or In mixture experiments, the factors are the components or ingredients of a mixture and consequently, their levels are not independent.

Let x_i represent the proportion of i^{th} component thus x_i being the proportion, we have the constraints:

$$0 \leq x_i \leq 1; i = 1,2 \ldots.$$

$$\& \sum_{i=1}^{q} x_i = 1$$

As a result, the factor space of q components reduce to a $q - 1$ - dimensional simplex

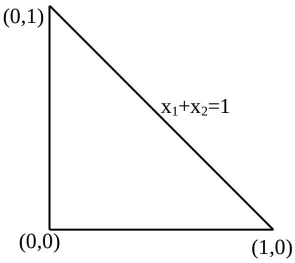
$s_{q-1} = \{\underline{x} = (x_1, x_2 x_3 \ldots. x_q / x_i \geq 0 \ and \ \sum_{i=1}^{q} x_i = 1\}$………1)

Another reason for blending different ingredients is to see if there exist some blends which gives more desirable product properties equation. In making different brands of detergents are available one may try blending 20 more brands for well effect on cloths.

Let $q = 2$

$$s_1 = \{\underline{x} = (x_1, x_2 x_3 \ldots. x_q / x_i \geq 0 \ and \ \sum_{i=1}^{q} x_i = 1\}$$

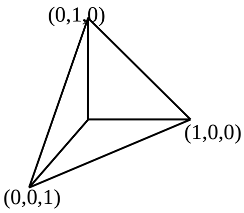
Graphically this can be represented as:



All the points lying on $x_1 + x_2 = 1$ are the points of the factor space (or experimental region)

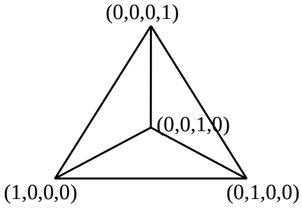
For $q = 2$; the factors space is straight line.

For $q = 3$;



It is an equilateral triangle (thus due to the constraint $\sum x_i = 1$,the dimension of factors space reduces by 1). We have $(q - 1)$ dimensional simplex, hence all the sides of the simplex will be equal.

For $q = 4$;



This is a tetrahedral. All the points lying on faces and inside, the tetrahedral constitute the points of experimental region. Scheff (1958,1963) introduced designs and models for mixture experiments. In 1958, he gave simplex lattice design and associate canonical polynomials. in 1963, he gave simplex centroid design and associated polynomials.

9.11 Simplex Lattice Design

Simplex lattice designs are characterized by symmetrical arrangement of points with in experimental region and a well-chosen polynomial equation to represent the response over the entire simplex region.

The polynomial has exactly, same number of points as there are points in the simplex lattice design. These designs consist of $\binom{q+m-1}{m}$ points, where each component proportions (i.e., x_i) can take (m+1) equally spaced value such as $x_i = 0,1/m, 2/m, 3/m \dots .1; i = 1,2 \dots . q$ remaining between $0 \; and \; 1$ and all possible mixture with these component proportions is used. This design is called $\{q, m\}$simplex lattice designs.

Example-1.: $\{3,3\}$ simplex lattice design

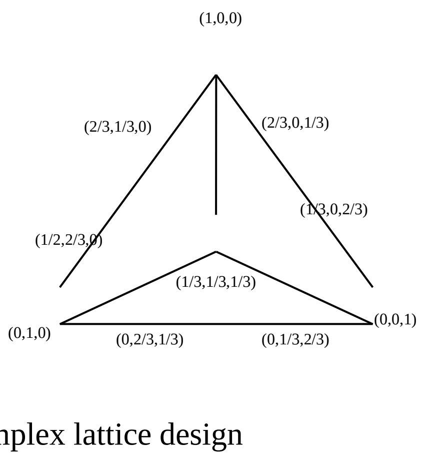
Here $m = 3$ & $q = 3$

Therefore, the number of points =$\binom{3+3-1}{3} = \binom{5}{3} = 10$

$x_i = 0,1/3,2/3,1 \; and \; i = 1,2,3$

The design pts will be: $(1,0,0), (0,1,0), (0,0,1), (1/3,2/3,0), (0,1/3,2/3), (1/3,0,2/3), (0,2/3,1/3), (2/3,1/3,0), (2/3,0,1/3), (1/3,1/3,1/3)$

.



Example-2.: $\{3,2\}$ simplex lattice design

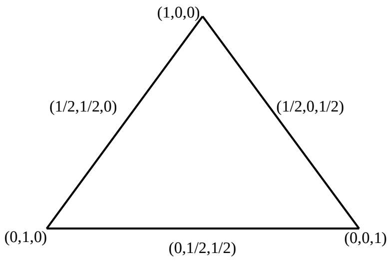
$q = 3, m = 2$

Here, number of design points $\binom{2+3-1}{2} = \binom{4}{2} = 6$

$x_i = 0,1/2,1 \; and \; i = 1,2,3$

The design points are

$(1,0,0), (0,1,0), (0,0,1), (0,1/2,1/2), (1/2,1/2,0), (1/2,0,1/2)$



The arrangement of points is symmetrical.

Canonical Polynomial

Scheffe also defined the canonical polynomials to be used with these simplex lattice design, also known as Scheffe form of polynomials. These are obtained by modifying the usual polynomial model in x_i by using the restriction.

The equation of an $m^{(th)}$ degree polynomial in general is:

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k + \ldots . \sum_{i_1<i_2<i_3 \ldots}^{q} \beta_{ij \ldots k} x_{i1} x_{i2 \ldots} x_{in} \quad (1)$$

The number of terms (parameters) in this (last) equation is $\binom{q+m}{m}$ .but for mixture experiments the terms in this polynomial have meaning for us only subject to restriction $x_1 + x_2 + x_3 \ldots\ldots + x_q = 1$.

We know that the parameters $\beta_i, \beta_{ij}, \beta_{ijk} \ldots \ldots$, associated with the terms are not unique. However, we may make the substitution; $x_q = 1 - \sum_{i=1}^{q-1} x_i$

For a q component mixture, an $m^{(th)}$ degree polynomial is known as {q, m} polynomial or canonical polynomial or canonical form of the polynomial .it has $\binom{q+m-1}{m}$ terms , this number is equal to the number of design points associated in {q, m} simplex lattice design.

A general equation for a 1st degree polynomial

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i \quad (2)$$

This equation can be rewritten as

$$\eta = \beta_0 \sum_{i=1}^{q} x_i + \sum_{i=1}^{q} \beta_i x_i$$

$$= \sum_{i=1}^{q} (\beta_0 + \beta_i) x_i \text{ Since } \sum_{i=1}^{q} x_i = 1$$

$$= \sum_{i=1}^{q} \beta_i^* x_i \quad (3)$$

$$\{ \beta_i^* = (\beta_0 + \beta_i)$$

The number of points in this model are polynomial is q which is same as the number of design points in {q,1} simple lattice design

A general equation for a 2nd degree polynomial is:

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_{ij} x_i\, x_j. \qquad (4)$$

$$\eta = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_i x_i \sum_{i<j}^{q} \beta_{ij} x_i x_j$$

Since $\sum_{i=1}^{q} x_i = 1, {x_i}^2 = x_i x_j = x_i(1 - \sum_{j \neq i}^{q} x_j)$

$$\eta = \beta_0 \sum x_i + \sum_i \beta_i x_i + \sum_j \beta_{ij} x_i (1 - \sum_{j \neq i} x_i) + \sum_{i<j} \beta_i x_i x_j$$

$$= \sum_{i=1}^{q} (\beta_0 + \beta_i + \beta_{ij}\,) x_i - \sum_{i=1}^{q} \beta_{ii} x_i \sum_{j \neq 1}^{q} x_i + \sum_{i<j}^{q} \beta_{ij} x_i x_j$$

$$\eta = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i<j}^{q} \beta_{iJ}^* x_i\, x_j. \qquad (5)$$

Where

$\beta_i^* = (\beta_0 + \beta_i + \beta_{ij})$ And $\beta_{ij}^* = (\beta_{ij} - \beta_{ii} + \beta_{jj})$

This is called {q,2} canonical polynomial or s- polynomial.

The number of terms in this model is: $q + \frac{q(q-1)}{2} = \frac{q(q+1)}{2}$ which are equal to the number of design points in {q,2} simplex lattice design.

The equation (5) can be written in the homogenous form we have

$$\eta = \sum_{i=1}^{q} \beta_i^* x_i \sum_{i=1}^{q-1} x_i + \sum_{i<j}^{q} \beta_{iJ}^* x_i\, x_j$$

$$= \sum_i \partial_{ii} x_i^2 + \sum_{i<j} \partial_j x_i x_j$$

$$= \sum \delta_{ij} x_i x_j$$

This is a *quadratic k model*.

Now, a 3rd degree polynomial is:

$$\eta = \beta_0 + \sum_{i}^{q} \beta_i x_i + \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{i \leq j \leq k}^{q} \beta_{ijx} x_i x_j\, x_k \qquad (7)$$

This can be rewritten as {q,3} canonical polynomial:

$$\eta = \sum_{i}^{q} \beta_i * x_i + \sum_{i<j}^{q} \beta_{iJ}^* x_i x_j + \sum_{i<j<k}^{q} \beta_{ijx} x_i x_j\, x_k \qquad (8)$$

This has $q(q+1)(q+2)/6$ points, same as the number of designs in $\{q, 3\}$ simplex lattice design.

This can be rewritten as a special cubic model as:

$$\eta = \sum_{i=1}^{q} \beta_i^* x_i + \sum_{i<j}^{q} \beta_{iJ}^* x_i\, x_j. + \sum_{i<j<k}^{q} \beta_{iJk}^* x_i\, x_j x_k \qquad (9)$$

This has $q(q^2+5)/6$, number of points $\binom{q+1}{1} + \binom{q}{3} = q(q^2+5)/6$

German and Hinman (1962) extended this work of canonical polynomial to quartic model while dealing with mixtures, we directly get β^*. Thus:

For {3,1} → $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3$(linear)

For {3,2} → $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$(quadratic)

For {3,3} → $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \partial_{12} x_1 x_2 (x_1 - x_2) + \partial_{13} x_1 x_3 (x_1 - x_3) + \partial_{23} x_2 x_3 (x_2 - x_3) + \beta_{123} x_1 x_2 x_3$ (full cubic)

And special cubic model is $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$

Any design with 7 or more design can be used to fit this model. Hence {3,3} can be used to fit these models.

There is 1-1 error between the number of design points of a {q, m} canonical polynomial. As a result, the points in the polynomial can be expressed as simple function of expected response at the points {q,m} simplex lattice design.

9.12 Self-Assessment Exercise

1. The following are the results of an experiment with cross-over design. Write the model, analyze and interpret the data.

Drug	*Person*				
	1	*3*	*5*	*7*	*9*
A	20.8	30.4	20.7	29.8	13.4
B	32.4	40.8	25.9	30.6	22.0

Drug	*Person*				
	2	*4*	*6*	*8*	*10*
B	50.2	40.4	60.8	75.2	30.2
A	56.2	48.3	59.8	70.4	50.4

2. What is a Response Surface Design? List any six desirable properties of a response surface design.
3. Discuss the fitting of first order response surface model in detail.
4. Explain the concept of blocking and orthogonality in response surface designs.

9.13 Summary

This unit provides an overview of various advance block designs such as Dual and linked block design, Lattice Designs, Cross-over designs, optimal designs, response surface design, weighing designs and mixture experiments.

9.14 References

- Gupta, V.K. & Nigam, A.K. (1978-79). *Handbook an analysis of Agriculture Experiment*, IASRI Publication.
- Montgomery, C.D. (1976). *Design and Analysis of Experiments*, Wiley, New York.
- Mukhopadhyay, P (2011). *Applied Statistics (2nd edition revised reprint), Books* & Allied (P) Ltd.
- Searle, S.R., Casella, G. and McCulloch, C.E. (1992): *Variance Components*, Wiley.

9.15 Further Reading

- Dean, A. & Voss, D. (1999). *Design and Analysis of Experiments*, Springer.
- Dey, A. (1986). *Theory of Block Designs*, Wiley Eastern.
- John P.W.M. (1971). *Statistical Design and Analysis of Experiments*, Mc Millan.
- Joshi, D.D. (1987). *Linear Estimation and Design of Experiments*, Wiley Eastern.
- Nigam, A, K., Puri, P.D., & Gupta, V.K. (1987-88). *Characterisation and Analysis of Block Design*, Wiley Eastern.
- Pearce, S.C. (1984). *Design of Experiments*, Wiley, New York.
- Rao, C.R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, North Holland.