
# BHARATHIDASAN UNIVERSITY

## Tiruchirappalli- 620024

## Tamil Nadu, India.

## Programme: M.Sc. Statistics

**Course Title:** Linear Models and Design of Experiments

**Course Code: 23ST01DEC**

## Unit-I

## Linear Models

**Dr. T. Jai Sankar**
**Associate Professor and Head**
**Department of Statistics**

**Ms. K. Nalini**
**Guest Faculty**
**Department of Statistics**

# UNIT-1

## LINEAR MODELS

Models generally represented specific components in a simplified manner to which further treatment and analysis may be carried out which leads to the final conclusion about the study. It is nothing but representation of a system. Mathematical models are in the form of functions, involving various parameters which represent the different components that are required under the specific study. The mathematical model either linear or non linear models.

**Classification of Linear Models:**

Classification of Linear Models are as follows:

- Probabilistic Model
- Deterministic Model

When the models are subject to the random fluctuations (variations) it is called probabilistic model. In this the linear model is highly applicable because it is simple to deal with the variable found a more realistic and appropriate for practical situations.

Example: $Y_i = aX_i + b$ is a linear model where a & b are constant. Parameters which is to be estimated using the set of sample observations $X_i$, $Y_i$ of size 'n'.

Suppose there are $n$ observations. In the linear model, we assume that these observation are the values taken by $n$ random variable $Y_1, Y_2, ....., Y_n$ satisfying the following conditions:

- $E(Y_i)$ is a linear combination of p unknown parameters $\beta_1, \beta_2, ....., \beta_p$,

  $E(Y_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + ..... + x_{ip}\beta_p, \qquad i = 1,2,.....n$

  where $x_{ij}$'s are known constants.

- $Y_1, Y_2, ....., Y_n$ are uncorrelated and normality distributed with variance $Var(Y_i) = \sigma^2$.

  The linear model can be rewritten by introducing independent normal random variable following $N(0, \sigma^2)$, as $Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ..... + x_{ip}\beta_p + \varepsilon_i, \qquad i = 1,2,.....n.$

  These equations can be written using the matrix notations as $Y = X\beta + \varepsilon$

where $Y$ is a $n \times 1$ vector observation, X is a $n \times p$ matrix of n observations on each of $X_1, X_2, ....., X_p$ variables, $\beta$ is a $p \times 1$ vector of parameters and $\varepsilon$ is a $n \times 1$ vector of random error components with $\varepsilon$ *follows* $N(0, \sigma^2)$, here $Y$ is called **study or dependent**. Variables $X_1, X_2, ....., X_p$ are called **explanatory or independent variables** and $\beta_1, \beta_2, ....., \beta_p$, are called as **regression coefficients**.

**Linear Model Assumption on Error Components**

**Least square estimation**

The least square estimate of $\beta$ is $Y = X\beta + \varepsilon$ is the value of $\beta$ which minimizes the error sum of squares $\varepsilon'\varepsilon$.

$$\text{Let } S = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = (Y'Y - 2\beta' X'Y + \beta' X' X\beta).$$

Minimizing S with respect to $\beta$ involves

$$\frac{\partial S}{\partial \beta} = 0 \text{ and } \Rightarrow X'X\beta = X'Y$$

which is termed as normal equation. This normal equation has a unique solution given by

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

Assuming rank (X)=p. Note that $\frac{\partial^2 S}{\partial \beta \partial \beta'} = X'X$ is a positive definite matrix. So $\hat{\beta} = (X'X)^{-1} X'Y$ is the value of $\beta$ which minimizes $\varepsilon'\varepsilon$ and is termed as ordinary least squares estimator of $\beta$.

- In this $\beta_1, \beta_2, ....., \beta_p$, are estimable and consequently, all the linear parametric function are estimable.
- $E(\hat{\beta}) = (X'X)^{-1} X'E(Y) = (X'X)^{-1} X'X\beta = \beta$
- $Var(\hat{\beta}) = (X'X)^{-1} X'Var(Y)X(X'X)^{-1} = \sigma^2 (X'X)^{-1}$
- If $\lambda'\hat{\beta}$ *and* $\mu'\hat{\beta}$ are the estimates of $\lambda'\beta$ *and* $\mu'\beta$ respectively, then
  - $Var(\lambda'\hat{\beta}) = \lambda' Var(\hat{\beta})\lambda = \sigma^2 (\lambda'(X'X)^{-1}\lambda)$
  - $Cov(\lambda'\hat{\beta}, \mu'\hat{\beta}) = \sigma^2 (\mu'(X'X)^{-1}\lambda)$
- $Y - X\hat{\beta}$ is called the residual vector
- $E(Y - X\hat{\beta}) = 0$

**Linear model with correlated observations:**

In the linear model $Y = X\beta + \varepsilon$ with $E(\varepsilon) = 0, Var(\varepsilon) = \sum$ *and* $\varepsilon$ is normally distributed, we find

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

$$Var(Y) = \sum$$

Assuming $\sum$ to be positive definite, so we can write $\sum P'P$

where P is a non-singular matrix. Pre-multiplying $Y = X\beta + \varepsilon$ by P, we get

$$PY = PX\beta + P\varepsilon \text{ (or)}$$

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

where

$Y^* = PY, X^* = PX$ *and* $\varepsilon^* = P\varepsilon$

Note that in this model $E(\varepsilon^*) = 0$ *and* $Var(\varepsilon*) = \sigma^2 I.$

**Distribution of $\alpha' Y$ :**

In the linear model $Y = X\beta + \varepsilon$, $\varepsilon$ follows $N(0, \sigma^2 I)$ consider a linear function $\alpha' Y$ which is normally distributed with

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

$$Var(\alpha' Y) = \sigma^2(\alpha'\alpha)$$

Then

$$\frac{\alpha' Y}{\sigma\sqrt{\alpha'\alpha}} \text{ follows } N\left(\frac{\alpha' X\beta}{\sigma\sqrt{\alpha'\alpha}}, 1\right)$$

Further, $\frac{(\alpha' Y)^2}{\sigma^2\sqrt{\alpha'\alpha}}$ has a non-central chi-square distribution with one degree of freedom and non-centrality parameter $\frac{(\alpha' X\beta)^2}{\sigma^2\sqrt{\alpha'\alpha}}$.

**Degrees of freedom**

A linear function $\alpha' Y$ of the observations $(\alpha \neq 0)$ is said to carry one degrees of freedom. A set of linear functions L'Y where L is r x n matrix, is said to have M degrees of freedom if there exist M linearly independent functions in the set and no more. Alternatively, the degrees of freedom carried by the set L'Y equals rank (L). When the set L'Y are the estimates of $\Lambda'\beta$, the degrees of freedom of the set 'L'Y will also be called the degrees of freedom for the estimates of $\Lambda'\beta$ .

**Sum of squares**

If $\alpha' Y$ is a linear function of observations, then the projection of Y on $\alpha'$ is the vector $\frac{Y'\alpha}{\alpha'\alpha}.\alpha.$ The square of this projection is called the sum of squares (SS) due to $\alpha' Y$ is given by $\frac{Y'\alpha}{\alpha'\alpha}$. Since $\alpha' Y$ has one degree of freedom, so the SS due $\alpha' Y$ to has one degree of freedom.

The sum of squares and the degrees of freedom arising out of the mutually orthogonal sets of functions can be added together to give the sum of squares and degrees of freedom for the set of all the function together and vice versa.

Let $X_1, X_2, ....., X_n$ has a multivariate normal distribution with mean vector $\mu$ and positive definite covariance matrix Σ. Let the two quadratic forms.

$X'A,X$ is distribution $\chi^2$ $with\, n_1$ degrees of freedom and non-centrality parameter $\mu' A_2\ \mu$. then $X'A_1\ X$ and $X'A_2\ X$ are iid $A_1 \sum A_2 = 0.$

**LINEAR AND ORTHOGONAL CONTRAST**

Two vector X and Y are said to be orthogonal if X'Y = Y'X = 0. The null vector is orthogonal to every vector X and is the only such vector.

**Linear Combination**

If $x_1, x_2, ....., x_m$ are m vector and $k_1, k_2, ....., k_m$ are m scalars, then $t = \sum_{i=1}^{m} k_i x_i$ is called the linear combination of $x_1, x_2, ....., x_m$.

**Linear Independence**

If $X_1, X_2, ....., X_m$ are m vector then they are said to be linearly independent if there exist scalars $k_1, k_2, ....., k_m$ such that $\sum_{i=1}^{m} k_i X_i = 0 \Rightarrow k_i = 0$ *for all* $i = 1,2,....., m.$ if there exist $k_1, k_2, ....., k_m$ with at least one $k_i$ to be nonzero, such that $\sum_{i=1}^{m} k_i x_i = 0$ than $x_1, x_2, ....., x_m$ are said to be **linearly dependent**.

Any set of vectors containing the null vector is linearly dependent.

Any set of non-null pair-wise orthogonal vectors is linearly independent.

If m > 1 vectors are linearly dependent, it is always possible to express at least one of them as a linear combination of the others.

**Linear function**

Let $K = (k_1, k_2, ....., k_m)'$ be a $m \times 1$vector of scalars and $X = x_1, x_2, ....., x_m$ be a $m \times 1$vector of variables, then $K'Y = \sum_{i=1}^{m} k_i y_i$ is called a linear function or linear form. The vector K is called the coefficient vector.

**Contrast**

The linear function $K'X = \sum_{i=1}^{m} k_i x_i$ is called a contrast in $x_1, x_2, ....., x_m$ if $\sum_{i=1}^{m} k_i = 0$.

For example, the linear functions $x_1 - x_2, 2x_1 - 3x_2 + x_3, \frac{x_1}{2} - x_2 + \frac{x_3}{3}$ are contrasts.

A linear function K'X is a contrast if it is orthogonal to a linear function $\sum_{i=1}^{m} x_i$ or to the linear function $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$.

Contrasts $x_1 - x_2, x_1 - x_3, ..., x_1 - x_j$ are linearly independent for all j = 1,2,3,……,m.

Every contrast in $x_1, x_2, ....., x_n$ can be written as a linear combination of (m-1) contrasts $x_1 - x_2, x_1 - x_3, ....., x_1 - x_m$.

**Gauss Markov Generalization**

If the linear parametric function $L'\beta$ is estimator $L'\hat{\beta}$ where $\hat{\beta}$ is a solution of $X'X\hat{\beta} = X'Y$ is the best linear unbiased estimator of $L'\beta$ in the sense of having minimum variance in the class of all linear and unbiased estimators of $L'\beta$.

**Estimator of $\sigma^2$ based on least squares estimation**

Consider an estimator of $\sigma^2$ as,

$$\hat{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})'(y - X\hat{\beta})$$

$$= \frac{1}{n-p}[y - X(X'X)^{-1}X'y]'[y - X(X'X)^{-1}X'y]$$

$$= \frac{1}{n-p}y'[I - X(X'X)^{-1}X'][I - X(X'X)^{-1}X']y$$

$$= \frac{1}{n-p}y'[I - X(X'X)^{-1}X']y$$

where the hat matrix $[I - X(X'X)^{-1}X']y$ is an idempotent matrix with its trace as

$$tr[I - X(X'X)^{-1}] = trI - trX(X'X)^{-1}X'$$

$$= n - trX(X'X)^{-1}X'X \text{ (using the result tr (AB) = tr (BA))}$$

$$= n - trI_p = n - p$$

Note that using $E(y'Ay) = tr(A\sum)$, *with* $Cov(y) = \sum$, *we have*

$$E(\hat{\sigma}^2) = \frac{\sigma^2}{n-p}tr[I - X(X'X)^{-1}X'] = \sigma^2$$

and so $\sigma^2$ is an unbiased estimator of $\sigma^2$.

**Estimator of $\beta$ based on least squares estimation:**

Let $y_1, y_2, ....., y_n$ be a sample of observations on $Y_1, Y_2, ....., Y_n$. the least-squares estimate of $\beta$ is the value $\hat{\beta}$ of $\beta$ for which the sum of squares due to errors, i.e.,

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) = y'y - 2X'y + \beta' X'X\beta$$

is minimum where $y = (y_1, y_2, ....., y_n)'$. differentiating $S^2$ with respect to $\beta$ and substituting it to be zero, the normal equations are obtained as

$$\frac{dS^2}{d\beta} = 2X'X\beta - 2X'y = 0 \ (or) \ X'X\beta = X'y$$

If X has full rank p, then (X'X) has a unique inverse and the unique least squares estimate of $\beta$ is

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

which is the best linear unbiased estimator of $\beta$ in the sense of having minimum variance in the class of linear and unbiased estimator. If the rank of X is not full, then generalized inverse is used for finding the inverse of (X'X).

If $L'\beta$ is a linear parametric function where $L = (\alpha_1, \alpha_{2,} ....., \alpha_p)'$ is a non-null vector, then the least squares estimate of $L'\beta$ is $L'\hat{\beta}$.

**Linear Estimation Functions**

A linear parametric function $\lambda'\beta$ of the parameter is said to be an estimable parametric function or estimable if there exists a linear function of random variables $\alpha' Y$ of Y where $Y = (Y_1, Y_2, ....., Y_n)'$ such that

$$E(\alpha' y) = \lambda'\beta$$

with $\alpha = (\alpha_1, \alpha_2, .....\alpha_n)'$ and $\lambda = (\lambda_1, \lambda_2, ....., \lambda_{,})'$ being vectors of known scalars.

**Gauss-Markov Theorem**

The Gauss-Markov theorem establishes that the generalized least-squares (GLS) estimator of $\beta$ given by $\hat{\beta} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}y,$ is Best Linear Unbiased Estimator (BLUE). By best $\beta$, we mean that $\hat{\beta}$ minimizes the variance for any linear combination of the estimated coefficients, $\alpha'\hat{\beta}$. We note that

$$E(\hat{\beta}) = [(X'\Omega^{-1}X)^{-1} X'\Omega^{-1}y] = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}E(y)$$
$$= (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}X\beta \ = \beta.$$

Thus $\hat{\beta}$ is an unbiased estimator of $\beta$.

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

$$\begin{aligned}Var(\hat{\beta}) &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]V(y)[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]' \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]\Omega[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}] \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}]\Omega[\Omega^{-1}X\,(X'\Omega^{-1}X)^{-1}] \\ &= (X'\Omega^{-1}X)^{-1}\end{aligned}$$

Thus,

$$Var(\alpha'\hat{\beta}) = \alpha' Var(\hat{\beta})\alpha = \alpha'[(X'\Omega^{-1}X)^{-1}]\alpha.$$

Let $\tilde{\beta}$ be another unbiased estimator of $\beta$ that is a linear combination of the data. Our goal, then, is to show that $Var(\alpha'\tilde{\beta}) \geq \alpha'(X\Omega^{-1}X)^{-1}\alpha$ with at least one $\alpha$ such that $Var(\alpha'\tilde{\beta}) \geq \alpha'(X'\Omega^{-1}X)^{-1}\alpha.$ we first note that we can write any other estimator of $\beta$ that is a linear combination of the data as

$$\tilde{\beta} = [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]y + b_0^*$$

where B is $p \times n$ matrix and $b_0^*$ is a $p \times 1$ vector of constants that appropriately adjusts the GLS estimator to from the alternative estimate. Then

$$\begin{aligned}E(\hat{\beta}) &= E([(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\,y - b_0) = ([(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\,E(y) - b_0^*) \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\,XB - b_0^* = [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta + BX\beta + b_0^* \\ &= \beta + BX\beta + b_0^*.\end{aligned}$$

Consequently, $\tilde{\beta}$ is unbiased if and only if both $b_0^* = 0$ and BX = 0. The covariance matrix of $\tilde{\beta}$ is

$$\begin{aligned}Var(\hat{\beta}) &= Var([(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\,y) \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]V(y)[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]' \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\Omega[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]' \\ &= [(X'\Omega^{-1}X)^{-1}X'\Omega^{-1} + B]\Omega[X\Omega^{-1}(X'\Omega^{-1}X)^{-1} + B'] \\ &= [(X'\Omega^{-1}X)^{-1} + B\Omega B']\end{aligned}$$

Because BX =0, which implies that (BX)'=X'B'=0. Then

$$\begin{aligned}Var(\alpha'\hat{\beta}) &= \alpha' Var(\hat{\beta})\alpha \\ &= \alpha'[(X'\Omega^{-1}X)^{-1} + B\Omega B']\alpha \\ &= \alpha'[(X'\Omega^{-1}X)^{-1}\alpha + \alpha' B\Omega B']\alpha \\ &= Var(\alpha'\hat{\beta}) + \alpha' B\Omega B'\alpha.\end{aligned}$$

We note that $\Omega$ is a positive definite matrix. Consequently, there exists some non-singular matrix K such that $\Omega$ = K'K as a result, $B\Omega B' = BK'KB'$ is at least positive semi definite matrix; hence $\alpha' B\Omega B'\alpha \geq 0.$ next note that we can define $\alpha^* = KB'\alpha.$ As a result,

$$\alpha' B\Omega B'\alpha = \alpha^*\alpha^* = \sum_{i=1}^{p}\alpha_i^{*2}$$

Which must be strictly greater than 0 for some $\alpha \neq 0$ unless B=0. Thus, the GLS estimate of $\beta$ is the best linear unbiased estimator.

**Test for Linear Hypothesis**

Let $Y_1, Y_2, ....., Y_n$ be a sequence of n independent random variables associated with responses. Then we can write it as

$$E(Y_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1,2,.....n, \quad j = 1,2,.....p$$

$$Var(Y_i) = \sigma^2$$

This is the linear model in the expectation from where $\beta_1, \beta_2, ....., \beta_p$, are the unknown parameters and $x_{ij}$ 's are the known values of independent covariates $X_1, X_2, ....., X_p$.

Alternatively, the linear model can be expressed as

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1,2,.....n, \quad j = 1,2,.....p$$

where $\varepsilon_i$, 's are identically and independently distributed random error component with mean 0 and variance $\sigma^2$, i.e.,

$$E(\varepsilon_i) = 0 \; Var(\varepsilon_i) = \sigma^2 \; and \; Cov(\varepsilon_i, \varepsilon_j) = 0 \; (i \neq j).$$

In matrix notations, the linear model can be expressed as

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

where $Y = (Y_1, Y_2, ....., Y_n)'$ *is a* $n \times 1$ vector of observations on the response variable,

The matrix $X = \begin{pmatrix} X_{11} \; X_{12} .....X_{1p} \\ X_{21} \; X_{22} .....X_{2p} \\ X_{31} \; X_{32} .....X_{3p} \\ \\ X_{n1} \; X_{n2} .....X_{np} \end{pmatrix}$ is a $n \times p$ matrix of n observation on p independent covariates $X_1, X_2, ....., X_p$,

$\beta = (\beta_1, \beta_2, ....., \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters (or regression coefficients) $\beta_1, \beta_2, ....., \beta_p$, associated with $X_1, X_2, ....., X_p$, respectively and $\varepsilon = (\varepsilon_1, \varepsilon_2, ....., \varepsilon_n)'$ is a $n \times 1$ vector of random errors or disturbances.

We assume that $E(\varepsilon) = 0$, the covariance matrix $V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_p, rank(X) = p$ in the context of analysis of variance and design of experiments, the matrix X is termed as the design matrix;

Unknown $\beta_1, \beta_2, ....., \beta_p$, are termed as effects.

The covariates $X_1, X_2, ....., X_p$, are counter variables or indicator variables where $x_{ij}$ counts the number of times the effect $\beta_j$ occurs in the i[th] observation $x_i$.

Note that in the linear regression model, the covariates are usually continuous variables.

When some of the covariates are counter variables, and is used in the analysis of covariance.

**Relationship between the Regression Model and Analysis of Variance Model**

The same linear model is used in the linear regression analysis as well as in the analysis of variance. So it is important to understand the role of a linear model in the context of linear regression analysis and analysis of variance.

Consider the multiple linear model

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + ..... + X_p\beta_p + \varepsilon.$$

In the case of analysis of variance model,

- the one-way classification considers only one covariate,
- two-way classification model considers two covariates,
- three-way classification model considers three covariates and so on.

If $\beta, \gamma$ and $\delta$ denote the effects associated with the covariates X, Z and W which are the counter variables, then in

One-Way model : $Y = \alpha + X\beta + \varepsilon$

Two-way model : $Y = \alpha + X\beta + Z\gamma + \varepsilon$

Three-way model : $Y = \alpha + X\beta + Z\gamma + W\delta + \varepsilon$ and so on.

**The regression parameters $\beta$'s can be fixed or random**

If all $\beta$'s are 's are unknown constants, they are called as parameters of the model and the model is β If all called as a fixed effect model or model I. The objective, in this case, is to make inferences .σabout the parameters and the error variance $\sigma^2$.

If for some *j*, $x_{ij} = 1$ for $i = 1,2,.....n$ then $\beta_j$ is termed an additive constant. In this case, $\beta_j$ occurs with every observation and so it is also called a general mean effect.

If all $\beta$'s are observable random variables expect the additive constant, then the linear model is termed as random effect model, model II or variance components model. The objective, in this case, is to make inferences about the variances of $\beta$'s i.e., $\sigma^2\beta_1, \sigma^2\beta_2 ....., \sigma^2\beta_p$ and error variance $\sigma^2$ and / or certain functions of them.

If some parameters are fixed and some are random variables, then the model is called a mixed effect model or model III. In the mixed effect model, at least one $\beta_j$ is constant and at least one $\beta_j$ is a random variable. The objective is to make inference about the fixed effect variance of random effects and error variance $\sigma^2$.

**Best Linear Unbiased Estimates (BLUE)**

The unbiased minimum variance linear estimate $\alpha' Y$ of an estimable function $\lambda'\beta$ is called the best linear unbiased estimate of $\lambda'\beta$.

- Suppose $\alpha'_1 Y$ *and* $\alpha'_2 Y$ are the BLUE of $\lambda_1'\beta$ *and* $\lambda_2'\beta$ respectively. Then $(a_1\,\alpha_1 + a_2\alpha_2)'Y$ is the BLUE of $(a_1\,\lambda_1 + a_2\lambda_2)'\beta$.
- If $\lambda'\beta$ is estimable, its best estimate is $\lambda'\hat{\beta}$ where $\hat{\beta}$ is any solution of the equations $X'X\beta = X'Y$.

**Design of Experiment**

Design of experiment means how to design an experiment in the sense that how the observations or measurements should be obtained to answer a query in a valid, efficient and economical way. The designing of experiment and the analysis of obtained data are inseparable. If the experiment is designed properly keeping in mind the question, then the data generated is valid and proper analysis of data provides the valid statistical inferences. If the experiment is not well designed, the validity of the statistical inferences is questionable and may be invalid. It is important to understand first the basic terminologies used in the experimental design.

**Experiment:** A way of getting an answer to a question which the experimenter wants to know. An experiment is a device or a means of getting an answer to the problem under consideration. Experiment can be classified into two categories;

i) **Absolute experiment:** Absolute experiments consist in determining the absolute value of some characteristics like,

a) Obtaining average intelligence quotient (I.Q) of a group of people.

b) Finding the correlation co-efficient between two variables in a bivariate distribution etc.

ii) **Comparative experiment:** Comparative experiments are designed to Compare the effect of two or more objects on some population characteristics.

Example: Comparison of different fertilizers.

Different kinds of verities of a crop.

Different cultivation processes etc.,

**Experimental unit:** For conducting an experiment, the experimental material is divided into smaller parts and each part is referred to as experimental unit. The experimental unit is randomly assigned to a treatment is the experimental unit. The phrase "randomly assigned" is very important in this definition. Example: i) In field experiments the plot of land is the experimental unit. In other experiments, unit may be a patient in a hospital, a lump of dough or a batch of seeds.

**Treatment:** Different objects or procedures which are to be compared in an experiment are called treatments. Example: In field experimentation different fertilizers or different varies of crop or different methods cultivation are the treatments.

**Blocks:** In agricultural experiments, most of the times we divide the whole experimental unit (field) into relatively homogeneous sub groups 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.

**Sampling unit:** The object that is measured in an experiment is called the sampling unit. This may be different from the experimental unit.

**Factor:** A factor is a variable defining a categorization. A factor can be fixed or random in nature. A factor is termed as fixed factor if all the levels of interest are included in the experiment. A factor is termed as random factor if all the levels of interest are not included in the experiment and those that are can be considered to be randomly chosen from all the levels of interest.

**Replication:** It is the repetition of the experimental situation by replicating the experimental unit. Replication means the execution of an treatments more than once. In other words, the repetition of treatments under investigation is known as replication.

**Experimental error:** The unexplained random part of variation in any experiment is termed as experimental error. An estimate of experimental error can be obtained by replication. Treatment design: A treatment design is the manner in which the levels of treatments are arranged in an experiment.

**Precision:** The reciprocal of the variance of the mean is termed as the precision. Thus for an experiment replicated r times is given by. $\frac{1}{Var\,(\bar{x})} = \frac{r}{\sigma^2}$ Where $\sigma^2$ is the error variance per unit.

Efficiency of a Design Consider the designs $D_1$ and $D_2$ with error variances per unit $\sigma_1^2$ and $\sigma_2^2$ and replications $r_1$ and $r_2$ respectively.

Then the variance of the difference between two treatment means is given by $\frac{2\sigma_1^2}{r_1}$ and $\frac{2\sigma_2^2}{r_2}$ for $D_1$ and $D_2$ respectively. Then the ratio $\frac{2\sigma_2^2}{r_2} \times \frac{r_1}{2\sigma_1^2}$ and $\frac{r_1}{\sigma_1^2} \div \frac{r_2}{\sigma_2^2}$ is termed as efficiency of design $D_1$ w.r.t. $D_2$.

**Uniformity Trials:** The fertility of the soil does not increase or decrease uniformity in any direction but is distributed over the entire field in an erratic manner. Uniformity trails enable us to have an idea about the fertility variation of the field. By uniformity trail, we mean a trail in which the field (experimental material) is divided into small units (plots) and the same treatment is applied on each of the units and their yields are recorded.

**Objectives of Design of Experiments**

One of the main objectives of designing an experiment is how to verify the hypothesis in an efficient and economical way. In the contest of the null hypothesis of equality of several means of normal populations having same variances, the analysis of variance technique can be used. Note that such techniques are based on certain statistical assumptions. If these assumptions are violated, the outcome of the test of hypothesis then may also be faulty and the analysis of data may be meaningless. So the main question is how to obtain the data such that the assumptions are met and the data is readily available for the application of tools like analysis of variance. The designing of such mechanism to obtain such data is achieved by the design of experiment. After obtaining the sufficient experimental unit, the treatments are allocated to the experimental units in a random fashion. Design of experiment provides a method by which the treatments are placed at random on the experimental units in such a way that the responses are estimated with the utmost precision possible.

**Principles of Experimental Design**

The purpose of designing an experiment is to increase the precision of the experiment. In order to increase the precision, we try to reduce the experimental error. For reducing the experimental error, we adopt certain techniques. These techniques form the form the basic principles of experimental designs. The basic principles of the experimental designs are replication, randomization and local control. The principles of experimental design;-

- Replication
- Local control
- Randomization

**Replication**

Replication means the repetition of the treatments under investigation. An experimenter resorts to replication in order to average out the influence of the chance factors on different experimental units. Thus, the repetition of treatment results is more reliable estimate than is possible with a single observation.

**Advantages of replication**

Replication serves to reduce experimental error and thus enables us to obtain more precise estimates of the treatment effects.

From statistical theory we know that the standard Error (S.E) of the mean of a sample size n is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population. Thus if a treatment is replicated r times, then the S.E of its mean effect is $\frac{\sigma}{\sqrt{n}}$ , where $\sigma^2$ is the variance of the individual plot is estimated from error variance. Thus, the precision of the experiment is inversely proportional to the square of the Replication has an important but limited role in increasing the efficiency of the design.

**Randomization**

We have seen that replication will provide an estimate of experimental error. For valid conclusions about our experimental results, we should have not merely an estimate of experimental error but it should be an unbiased estimate. Also, if our conclusions are to be valid, the treatment means and also differences among treatment means should be estimated without any bias. For the purpose we use the technique of randomization. When all the treatments have equal chances of being allocated to different experimental units it is known as randomization. The following are the main objectives of randomization.

- The validity of the statistical test of the Significance. i.e.) t-test for testing the significance of the difference of two means. F-test for testing the homogeneity of variance.
- The purpose of randomness is to assure that the source of variation, not controlled in the experiment operate randomly. Randomization eliminates bias in any form.

**Local control**

We know that the estimate of experimental error is based on the variations from experimental unit to experimental unit. In other words, the error in an experiment is a measure of "within block" variation. This suggests that if we group the homogeneous experimental units into blocks, the experimental error will be reduced considerably. If the experimental material, say field for agriculture experimentation is heterogeneous and different treatment are allocated to various units at random over the entire field the soil heterogeneous will also enter the uncontrolled factors and thus increase the experimented error. It is desirable to reduce the experimental error as for as practicable without unduly increasing the number of replications, so that even smaller difference between treatments can be detected as significant. The process of reducing the experimental error by dividing relatively heterogeneous experimental area (field) into homogeneous blocks is known as local control.

## Analysis of Variance (ANOVA)

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

## Completed Randomized Design (CRD)

In this design the experimental units are allotted at random to the treatments, so that every unit gets the same chance of receiving every treatment.

For example: 25 Let there be five treatments each to be replicated four times. There are, therefore, 20 plots. Let these plots be numbered from 1 to 20 conveniently.

When a coin is tossed, there are two events, that is, either the head comes up, or the tail. We denote the “head” by H and the “tail” by T.

## Layout of CRD

| | | | |
|---|---|---|---|
| **1**<br>A | **2**<br>C | **3**<br>A | **4**<br>D |
| **5**<br>B | **6**<br>D | **7**<br>B | **8**<br>D |
| **9**<br>C | **10**<br>B | **11**<br>C | **12**<br>D |
| **13**<br>B | **14**<br>D | **15**<br>A | **16**<br>C |

## Advantages of CRD

- It is easy to layout the design.
- It results in the maximum use of the experimental units since all the experimental materials can be used.
- It allows complete flexibility as any number of treatments and replicates may be used. The number of replicates , if desired, can be varied from treatment to treatment.
- The statistical analysis is easy even if the number of replicates are not the same for all treatments.
- It provides the maximum number of degrees of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small experiments.

**Disadvantages of CRD**

- In certain circumstances, the design suffers from the disadvantage of being inherently less informative than other more sophisticated layouts. This usually happens if the experimental material is not homogeneous.
- Since, randomisation is not restricted in any direction to ensure that the units receiving one treatment are similar to those of receiving other treatment, the whole variations among the experimental units is included in the residual variance.
- This makes the design less efficient and results in less sensitivity in detecting significant effects.

**Applications:** CRD is most useful in laboratory technique and methodological studies, e.g., in physics, chemistry, in chemical and biological experiments , in some green house studies, etc.

**Statistical Analysis of CRD:** The model is

$$y_{ij} = \mu + t_i + e_{ij}$$

$$\text{for all } \; i = 1,2,3.....,k \; and \; j = 1,2,....n_i$$

where $y_{ij}$ is the yield

$\mu$ is the general mean effect

$t_i$ is the general mean effect

$e_{ij}$ is the error term mean zero and variance $\sigma^2$

$E(y_{ij}) = \mu + t_i, \; i = 1,2,.....k$ can be estimated by method of least square that is minimizing error sum of square

$$E\,(e'e)^2 = \sum_{ij}(y_{ij} - E(y_{ij})^2) = \sum_{ij}(y_{ij} - (\mu + t_i))^2 \; \frac{\partial(e'e)}{\partial\mu} = 0$$

$$2\sum_{ij}(y_{ij} - \mu - t_i)\,(1) = 0$$

$$-2\sum_{ij}(y_{ij} - \mu - t_i) \; = 0$$

$$\sum_{ij}(y_{ij} - \mu - t_i) \; = \frac{0}{-2} = 0$$

where $\sum_i y_{ij} \; = G$, G = Grand Total

$$G = \sum_i n_i \; \mu + \sum_i n_i \; t_i \quad .....(1)$$

$$\frac{\partial(e'e)}{\partial t_i} = 0$$

$$2\sum_j(y_{ij} - \mu - t_i)\,(-1) = 0 - 2\sum_j(y_{ij} - \mu - t_i) = 0$$

$$\sum_j(y_{ij} - \mu - t_i) = \frac{0}{-2} = 0$$

$$\sum_j y_{ij} - \sum_j \mu - \sum_j t_i = 0$$

where $\sum_j y_{ij} = T_i$

$$T_i = n_i\mu + n_i t_i = 0 \quad \text{.....(2)}$$

From equation (1)

$$\sum_i n_i\, t_i = 0 \quad \text{and} \quad \sum_i n_i = n$$

$$G = n\hat{\mu} + 0$$

$$\frac{G}{n} = \hat{\mu}$$

From equation (2)

$$T_i - n_i\hat{\mu} = n_i t_i$$

$$\frac{T_i}{n_i} - \frac{n_i G}{n_i n} = \hat{t}_i$$

$$\frac{T_i}{n_i} - \frac{G}{n} = \hat{t}_i$$

Error sum of squares

$$E(e'e)^2 = 0\sum_{ij}(y_{ij} - \mu - t_i)^2$$

$$\sum_{ij}(y_{ij} - \mu - t_i)\,(y_{ij} - \mu - t_i)$$

$$\sum_{ij} y_{ij}(y_{ij} - \mu - t_i) + \text{ other terms are vanished.}$$

$$\sum_{ij}[y_{ij}{}^2 - \hat{\mu}\, y_{ij} - \hat{t}_i y_{ij}] = \sum_{ij} y_{ij}{}^2 - \hat{\mu}\sum_{ij} y_{ij} - \hat{t}_i \sum_{ij} y_{ij}$$

$$= \sum_{ij} y_{ij}{}^2 - \hat{\mu}\sum_{ij} y_{ij} - \sum_i \hat{t}_i \sum_{ij} y_{ij}$$

$$= \sum_{ij} y_{ij}{}^2 - \frac{G}{n}G - \sum_i\left(\frac{T_i}{n_i} - \frac{T_i G}{n}\right)$$

$$= \sum_{ij} y_{ij}{}^2 - \frac{G^2}{n} - \left(\sum_i \frac{T_i^2}{n_i} - \frac{G^2}{n}\right)$$

where $\sum_j y_{ij} = T_i$

Error Sum of Square (E.S.S) = Total Sum of Square (T.S.S) - Treatment Sum of Square (Tr.S.S)

where $\frac{G^2}{n}$ is the Correction Factor.

**ANOVA for CRD**

| Source of Variation | d.f | Sum of Square (S.S) | Mean Sum of Square (M.S.S) | F-Ratio |
|---|---|---|---|---|
| **Treatments** | k-1 | Tr.S.S= $\sum_i \frac{T_i^2}{n_i} - \frac{G^2}{n}$ | $M.S.S.T = \frac{Tr.S.S}{k-1}$ | $F = \frac{M.S.S.T}{M.S.S.E}$ |
| **Error** | n-k | By Subtraction<br>E.S.S = T.S.S - Tr.S.S | $M.S.S.E = \frac{Er.S.S}{n-k}$ | |

Under the null hypothesis, $H_0=t_1=t_2=\ldots\ldots t_k$ against the alternative that all t's are not equal, the test Statistics $F = \frac{M.S.S.T}{M.S.S.E}$ *follows* $F_{(k-1,n-k)}$

i.e., F follows F (Central) distribution with $(k-1, n-k)$ d.f.

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. Otherwise $H_0$ accepted.

**Randomised Block Design (RBD)**

If all the treatments are applied at random relatively homogeneous units within each strata or block and replicated over all the blocks. The design is a randomised block design.

**Advantages of RBD**

- **Accuracy:** This design has been shown to be more efficient or accurate than C.R.D for most types of experimental work. The elimination of between S.S. from residual S.S. usually results in a decrease of error mean S.S.
- **Flexibility:** In R.B.D no restriction are placed on the number of treatments or the number of replicates. In general, at least two replicates are required to carry out the test of significance (factorial design is an exception). In addition, control (check) or some other treatments may be included more than once without complications in the analysis.
- **Ease of Analysis:** Statistical analysis is simple and rapid. More-over the error of any treatment can be isolated and any number of treatments may be omitted from the analysis without complicating it.

**Disadvantages of RBD**

- RBD may give misleading results if blocks are not homogeneous.
- RBD is not suitable for large number of treatments in that case the block size will increase and it may not be possible to keep large blocks homogeneous.
- If the data on more than two plots is missing, the statistical analysis becomes quite tedious and complicated.

**Layout of RBD**

Let us consider five Treatments A, B, C, D, E each replicated 4 times we divided the whole experimental area into 4 relatively homogeneous block and each in to 5 units the treatments allocated at random to the blocks particular layout may be follows:

| | | | | | |
|---|---|---|---|---|---|
| **Block 1** | A | B | C | D | E |
| **Block 2** | B | C | D | E | A |
| **Block 3** | C | D | E | A | B |
| **Block 4** | D | E | A | B | C |

| | | | | | means | total |
|---|---|---|---|---|---|---|
| | $y_{11}$ | $y_{12}$ | .... | $y_{1r}$ | $\bar{y}_{1.}$ | $T_{1.}$ |
| | $y_{21}$ | $y_{22}$ | .... | $y_{2r}$ | $\bar{y}_{2.}$ | $T_{2.}$ |
| | .. | ... | ... | ... | ... | .. |
| | $y_{i1}$ | $y_{i2}$ | .... | $y_{ij}y_{ir}$ | $\bar{y}_{i.}$ | $T_{i.}$ |
| | ... | ... | ... | ... | .... | .... |
| | ... | ... | ... | .... | .... | ..... |
| | $y_{t1}$ | $y_{t2}$ | .... | $y_{tr}$ | $\bar{y}_{t.}$ | $T_{t.}$ |
| *means* | $\bar{y}_{.1}$ | $\bar{y}_{.2}$ | .... | $\bar{y}_{.r}$ | | ↓ |
| *total* | $T_{.1}$ | $T_{.2}$ | .... | $T_{.r}$ | → | $G$ |

**Statistical Analysis of RBD**

The model is

$$y_{ij} = \mu + t_i + b_j + e_{ij} \quad for\ all\ \ i = 1,2,3.....,t\ and\ \ j = 1,2,....r$$

where $y_{ij}$ is the response or the yield of the experimental unit receiving the $i^{th}$ treatment in the $j^{th}$ block;

$\mu$ is the general mean effect

$t_i$ is the effect due to the $i^{th}$ treatment

$b_j$ is the effect due to $j^{th}$ block or replicate

$\mu$ is the general mean effect

$t_i$ is the general mean effect

$e_{ij} \underset{follows}{i.i.d}\ N(0, \sigma_e^2)$

where $\mu$, $t_i$ and $b_j$ are constants so that $\sum_{i=1}^{t} t_i = 0\ and\ \sum_{j=1}^{r} b_j = 0$

If we write $= \sum_i \sum_j y_{ij} = G = Grand\ Total$

$$\sum_j y_{ij} = T_i = Total\ \ for\ i^{th}\ \ treatment$$

$$\sum_i y_{ij} = B_j = Total\ \ for\ \ j^{th}\ \ Block$$

where $\mu$, $t_i$ and $b_j$ are estimated by the method of least squares

$$E = \sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (y_{ij} - \mu - t_i - b_j)^2 \quad .....(1)$$

Differentiate with respect to $\mu$

$$\frac{\partial E}{\partial \mu} = 0$$

$$2\sum_i \sum_j (y_{ij} - \mu - t_i - b_j)\,(1) = 0$$

$$-2\sum_i \sum_j (y_{ij} - \mu - t_i - b_j) = 0$$

$$\sum_i \sum_j (y_{ij} - \mu - t_i - b_j) = 0 = \frac{0}{-2} = 0$$

$$\sum_i \sum_j y_{ij} - \sum_i \sum_j \mu - \sum_i \sum_j t_i - \sum_i \sum_j b_j = 0$$

$$\sum_i \sum_j y_{ij} - tr\mu - r\sum_i t_i - t\sum_j b_j) = 0$$

where $\sum_i \sum_j y_{ij} = G$

$$G - tr\mu - r\sum_i t_i - t\sum_j b_j = 0 \text{ ..... (2)}$$

Differentiate with respect to $t_i$

$$\frac{\partial E}{\partial t_i} = 0$$

$$2\sum_j (y_{ij} - \mu - t_i - b_j)(-1) = 0 - 2\sum_j (y_{ij} - \mu - t_i - b_j) = 0$$

$$\sum_j (y_{ij} - \mu - t_i - b_j) = \frac{0}{-2} = 0$$

$$\sum_j y_{ij} - \sum_j \mu - \sum_j t_i - \sum_j b_j = 0$$

where $\sum_j y_{ij} = T_i$

$$T_i = r\mu + rt_i - \sum_j b_j = 0 \text{ .....(3)}$$

Differentiate with respect to $t_i$

$$\frac{\partial E}{\partial b_j} = 0$$

$$2\sum_i (y_{ij} - \mu - t_i - b_j)(-1) = 0 - 2\sum_i (y_{ij} - \mu - t_i - b_j) = 0$$

$$\sum_i (y_{ij} - \mu - t_i - b_j) = \frac{0}{-2} = 0$$

$$\sum_i y_{ij} - \sum_i \mu - \sum_i t_i - \sum_i b_j = 0$$

where $\sum_i y_{ij} = B_j$

$$B_i = t\mu + \sum_i t_i + tb_j = 0 \text{ .....(4)}$$

$$\sum_{i=1}^{t} t_i = 0 \text{ and } \sum_{j=1}^{r} b_j = 0$$

From equation (2)

$$G = tr\mu$$

$$\frac{G}{tr} = \hat{\mu}$$

From equation (3)

$$T_i = r\hat{\mu} + rt_i$$

$$T_i - r\hat{\mu} = rt_i$$

$$\frac{T_i}{r} - \frac{G}{tr} = \hat{t}_i$$

From equation (4)

$$B_j - t\hat{\mu} = tb_j$$

$$\frac{B_j}{t} - \frac{G}{tr} = \hat{b}_j$$

$$T_i - r\hat{\mu} = rt_i$$

$$\frac{T_i}{r} - \frac{G}{tr} = \hat{b}_j$$

Error sum of squares

$$E = \sum_i \sum_j (y_{ij} - \mu - t_i - b_j)^2$$

$$E = \sum_i \sum_j (y_{ij} - \mu - t_i - b_j)(y_{ij} - \mu - t_i - b_j)$$

$$E = \sum_i \sum_j y_{ij}(y_{ij} - \mu - t_i - b_j) + otherterms\ are\ vanished$$

$$E = \sum_i \sum_j y_{ij}^2 - \hat{\mu} \sum_i \sum_j y_{ij} - \sum_i \sum_j y_{ij}\hat{t}_i - \sum_i \sum_j y_{ij}\hat{b}_j$$

$$= \sum_i \sum_j y_{ij}^2 - \frac{G}{tr}G - \sum_i \sum_j y_{ij}\left(\frac{T_i}{r} - \frac{G}{tr}\right) - \sum_i \sum_j y_{ij}\left(\frac{B_j}{r} - \frac{G}{tr}\right)$$

where $\sum_i \sum_j y_{ij} = G; \sum_j y_{ij} = T_i; \sum_i y_{ij} = B_j$

$$= \sum_i \sum_j y_{ij}^2 - \frac{G^2}{n} - \left(\frac{\sum_i T_i^2}{r} - \frac{G^2}{tr}\right) - \left(\frac{\sum_j B_j^2}{t} - \frac{G^2}{tr}\right)$$

Error Sum of Square (E.S.S) = Total Sum of Square (T.S.S) - Treatment Sum of Square (Tr.S.S) - Block Sum of Square (B.S.S)

where $\frac{G^2}{tr}$ is the Correction Factor.

Total Sum of Square $= \sum_i \sum_j y_{ij}^2 - \frac{G^2}{tr}$

Treatment Sum of Square $= ST^2 = \frac{\sum_i T_i^2}{r} - \frac{G^2}{tr}$

Block Sum of Square $= SB^2 = \frac{\sum_j B_j^2}{t} - \frac{G^2}{tr}$

**ANOVA for RBD**

| Source of Variation | d.f. | Sum of Square (S.S) | Mean Sum of Square (M.S.S) | F-Ratio |
|---|---|---|---|---|
| **Treatments** | t-1 | $ST^2$ | $ST^2 = ST^2/t-1$ | $FT = ST^2/SE^2$ |
| **Block or replicates** | r-1 | $SB^2$ | $SB^2 = SB^2/r-1$ | $FB = SB^2/SE^2$ |
| **Error** | (t-1) (r-1) | $SE^2$ | $SE^2 = SE^2/(t-1)(r-1)$ | |
| **Total** | N-1 | | | |

Under the null hypothesis, $H_0=t_1=t_2=\ldots\ldots t_k$ against the alternative that all t's are not equal, the test Statistics $F_T = \frac{S_T^2}{S_E^2}$ *follows* $F_{((t-1),(t-1)(r-1))}$

i.e., F follows F (Central) distribution with (k-1, n-k) d.f.

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. Otherwise $H_0$ accepted.

i.e., $F_T$ follows F(central) distribution with [(t-1), (t-1)(r-1)] d.f. Thus if $F_T$ is greater than tabulated F for [(t-1), (t-1)(r-1)] d.f. at certain level of significance, usually 5 % then we reject the null hypothesis $H_{0t}$ and conclude that the treatments differ significantly. If $F_t$ is less than tabulated value then $F_T$ is not significant and we conclude that the data do not provide any evidence against the null hypothesis which may be accepted. Similarly under the null hypothesis $H_{0b}=b_1=b_2=\ldots=b_r$, against the alternative that b" s are not equal, the test statistics is: $F_b = \frac{S_B^2}{S_E^2}$ *follows* $F_{((t-1),(t-1)(r-1))}$

**Latin Square Design (LSD)**

LSD is defined for eliminating the variation of two factors called row and column in this design. The number of treatments is equal to the number of replications. Layout of design In this design the number of treatments is equal to the number of replications.

In this case of m treatments there have to be m×m = m² experimental units. The whole of the experimental area is divided into m² experimental units (plots) arranged in a square so that each row as well each column contain m units.

The m treatments are allocated at random to these rows and columns in such a way that every treatment occurs only once in each row and in each column. Such a layout is LSD.

$2\times 2$ *Layouts*

| | |
|---|---|
| A | B |
| B | A |

$3\times 3$ *Layouts*

| | | |
|---|---|---|
| A | B | C |
| B | C | A |
| C | A | B |

$4\times 4$ *Layouts*

| | | | |
|---|---|---|---|
| A | B | C | D |
| B | C | D | A |
| C | D | A | B |
| D | A | B | C |

$5 \times 5$ *Layouts*

| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| B | C | D | E | A |
| C | D | E | A | B |
| D | E | A | B | C |
| E | A | B | C | D |

**Example:** An animal feeding experiment where the column groups may corresponding with initial weight and the row group with age.

**Standard Latin square:** A Latin in which the treatments say A, B, C etc occur in the first row and first column in alphabetical order is called standard Latin square.

**Example:**

| | |
|---|---|
| A | B |
| B | A |

**Advantages of LSD**

- With two way grouping LSD controls more of the variation than CRD or RBD.
- The two way elimination of variation as a result of cross grouping often results in small error mean sum of squares.
- LSD is an incomplete 3-way layout. Its advantage over the complete 3-way layout is that instead of $m^3$ experimental units only $m^2$ units are needed. Thus, a 4x4 LSD results in saving of $m^3 = 4^3 - 4^2 = 64 - 16 = 48$ observations over a complete 3-way layout.
- The statistical analysis is simple though slightly complicated than for RBD. Even 1 or 2 missing observations the analysis remains relatively simple.
- More than one factor can be investigated simultaneously.

**Disadvantages of LSD**

- LSD is suitable for the number of treatments between 5 and 10 and for more than 10 to 12 treatments the design is seldom used. Since in that case, the square becomes too large and does not remain homogeneous.
- In case of missing plots the statistical analysis becomes quite complex.
- If one or two blocks in a field are affected by some disease or pest. We can't omit because the number of rows columns and treatments have to be equal.

**Statistical Analysis of LSD**

Let $y_{ijk}$ $(i, j, k = 1,2,.....m)$ denote the response from the unit in the i[th] row j[th] column and receiving the k[th] treatment. The model is

$$y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk}; \qquad i, j, k = 1,2,...., m$$

where $\mu$ is the constant mean effect; $r_i$, $c_j$ and $t_k$ due to the i[th] row j[th] column and k[th] treatment respectively and $e_{ijk}$ is error effect due to random component assumed to be normally distributed with mean zero and variance $\sigma_e^2$ $i.e., e_{ijk}$ $follow$ $N(0, \sigma_e^2)$

If we write

G = total of all the m observations
$R_i$ = total of the m observations in the i[th] row
$C_j$ = total of the m observations in the j[th] column
$T_k$ = total of the m observations from k[th] treatment

Estimation by the method of least squares

$$E(e_{ijk})^2 = \sum_{ijk}(y_{ijk} - \mu - r_i - c_j - t_k)^2 \ldots\ldots\ldots\ldots\ldots (1)$$

$$\frac{\partial E}{\partial \mu} = 0, \frac{\partial E}{\partial r_i} = 0, \frac{\partial E}{\partial c_j} = 0, \frac{\partial E}{\partial t_k} = 0$$

Differentiate with respect to $\mu$ in equation (2)

$$\frac{\partial E}{\partial \mu} = 2\sum_{ijk}(y_{ijk} - \mu - r_i - c_j - t_k)(-1) = 0$$

$$\sum_{ijk} y_{ijk} - \sum_{ijk}\mu - \sum_{ijk} r_i - \sum_{ijk} c_j - \sum_{ijk} t_k = 0$$

where $\sum_{ijk} y_{ijk} = G,\ \ I, j, k = m^2, I, j = m, I, k = m$

$$G - m^2\mu - m\sum_i r_i - m\sum_j c_j - m\sum_k t_k = 0 \ldots\ldots\ldots\ldots\ldots (2)$$

Differentiate with respect to $r_i$ in equation (3)

$$\frac{\partial E}{\partial \mu} = 2\sum_{jk}(y_{ijk} - \mu - r_i - c_j - t_k)(-1) = 0$$

$$\sum_{jk} y_{ijk} - \sum_{jk}\mu - \sum_{jk} r_i - \sum_{jk} c_j - \sum_{jk} t_k = 0$$

where $\sum_{jk} y_{ijk} = R_i,\ \ I, j, k = m^2,\ \ I, j = m, I, k = m$

$$R_i - m\mu - mr_i - m\sum_j c_j - m\sum_k t_k = 0 \ldots\ldots\ldots\ldots\ldots (3)$$

Differentiate with respect to $c_j$ in equation (4)

$$\frac{\partial E}{\partial c_j} = 2\sum_{ik}(y_{ijk} - \mu - r_i - c_j - t_k)(-1) = 0$$

$$\sum_{ik} y_{ijk} - \sum_{ik}\mu - \sum_{ik} r_i - \sum_{ik} c_j - \sum_{ik} t_k = 0$$

where $\sum_{ik} y_{ijk} = C_j,\ i, j, k = m^2,\ i, j = m,\ i, k = m$

$$C_j - m\mu - m\sum_i r_i - mc_j - m\sum_k t_k = 0 \ldots\ldots\ldots\ldots\ldots (4)$$

Differentiate with respect to $t_k$ in equation (5)

$$\frac{\partial E}{\partial t_k} = 2\sum_{ij}(y_{ijk} - \mu - r_i - c_j - t_k)(-1) = 0$$

$$\sum_{ij} y_{ijk} - \sum_{ij}\mu - \sum_{ij} r_i - \sum_{ij} c_j - \sum_{ij} t_k = 0$$

where $\sum_{ij} y_{ijk} = T_k,\ i, j, k = m^2,\ i, j = m,\ i, k = m$

$$T_k - m\mu - m\sum_i r_i - mc_j - mt_k = 0 \ldots\ldots\ldots\ldots\ldots (5)$$

The equation (2), (3), (4) and (5) are not independent

We assume that,

$$\sum_i r_i = 0, \sum_j c_j = 0 \ \ and \ \sum_k t_k = 0$$

From equation (2)

$$G = m^2\mu$$

$$\frac{G}{m^2} = \hat{\mu}$$

From equation (3)

$$R_i - m\mu - mr_i = 0$$

$$R_i - m\hat{\mu} = mr_i$$

$$\frac{R_i}{m} - \frac{mG}{mm^2} = \hat{r}_i$$

$$\frac{R_i}{m} - \frac{G}{m^2} = \hat{r}_i$$

From equation (4)

$$C_j - m\mu - mc_j = 0$$

$$C_j - m\hat{\mu} = mc_j$$

$$\frac{C_j}{m} - \frac{mG}{mm^2} = \hat{c}_j$$

$$\frac{C_j}{m} - \frac{G}{m^2} = \hat{c}_j$$

From equation (5)

$$T_k - m\mu - mt_k = 0$$

$$T_k - m\hat{\mu} = mt_k$$

$$\frac{T_k}{m} - \frac{mG}{mm^2} = \hat{t}_k$$

$$\frac{T_k}{m} - \frac{G}{m^2} = \hat{t}_k$$

Error Sum of Square

$$E(e_{ijk})^2 = \sum_{ijk}(y_{ijk} - \mu - r_i - c_j - t_k)^2$$

$$= \sum_{ijk}(y_{ijk} - \mu - r_i - c_j - t_k)(y_{ijk} - \mu - r_i - c_j - t_k)$$

$$= \sum_{ijk}(y_{ijk} - \mu - r_i - c_j - t_k)(y_{ijk}) + \textit{other terms are vanished}$$

$$= \sum_{ijk} y_{ijk}^2 - \hat{\mu}\sum_{ijk} y_{ijk} - \sum_{ijk} y_{ijk}\hat{r}_i - \sum_{ijk} y_{ijk}\hat{c}_j - \sum_{ijk} y_{ijk}\hat{t}_k)$$

$$= \sum_{ijk} y_{ijk}^2 - \frac{G}{m^2}\sum_{ijk} y_{ijk} - \sum_{ijk} y_{ijk}\left(\frac{R_i}{m} - \frac{G}{m^2}\right) - \sum_{ijk} y_{ijk}\left(\frac{C_j}{m} - \frac{G}{m^2}\right)_i - \sum_{ijk} y_{ijk}\left(\frac{T_k}{m} - \frac{G}{m^2}\right)$$

$$= (\sum_{ijk} y_{ijk}^2 - \frac{G}{m^2}) - \left(\frac{\sum_i R_i^2}{m} - \frac{G^2}{m^2}\right) - \left(\frac{\sum_j C_j^2}{m} - \frac{G^2}{m^2}\right) - \left(\frac{\sum_k T_k^2}{m} - \frac{G^2}{m^2}\right)$$

Total Sum of Square $= \sum_{ijk} y_{ijk}^2 - \frac{G^2}{m^2}$

Row Sum of Square $= S_R^2 = \frac{\sum_i R_i^2}{m} - \frac{G^2}{m^2}$

Column Sum of Square $= S_C^2 = \frac{\sum_j C_j^2}{m} - \frac{G^2}{m^2}$

Treatment Sum of Square $= S_T^2 = \frac{\sum_k T_k^2}{m} - \frac{G^2}{m^2}$

Error Sum of Square $= S_E^2 = \sum_{ijk} y_{ijk}^2 - \frac{\sum_i R_i^2}{m} - \frac{\sum_j C_j^2}{m} - \frac{\sum_k T_k^2}{m}$

**ANOVA Table for LSD**

| Source of Variation | d.f | Sum of Square (S.S) | Mean Sum of Square (M.S.S) | F-Ratio |
|---|---|---|---|---|
| **Row** | m-1 | $S_R^2$ | $S_R^2 = S_R^2/(m-1)$ | $F_R = S_R^2/S_E^2$ |
| **Column** | m-1 | $S_C^2$ | $S_C^2 = S_C^2/(m-1)$ | $F_C = S_C^2/S_E^2$ |
| **Treatment** | m-1 | $S_T^2$ | $S_T^2 = S_T^2/(m-1)$ | $F_T = S_T^2/S_E^2$ |
| **Error** | (m-1)(m-2) | $S_E^2$ | $S_E^2 = S_E^2/(m-1)(m-2)$ | |
| **Total** | $m^2$-1 | | | |

**Let us set up null hypothesis**

For row effects $H_{0r}=r_1=r_2=\ldots=r_m=0$

For column effects $H_{0c}=c_1=c_2=\ldots=c_m=0$

For treatment effects $H_{0t}=t_1=t_2=\ldots=t_m=0$

**Alternative Hypotheses**

For row effects, $H_{1r}$: At least two $r_i$'s are different

For column effects, $H_{1c}$: At least two $c_i$'s are different

For treatment effects, $H_{1t}$: At least two $t_i$'s are different

d.f. under the null hypotheses $H_r$, $H_b$ and $H_t$, respectively.

Let $F_\alpha = F_\alpha\{(m-1), (m-1)(m-2)\}$ be tabulated value of F for [(m-1),(m-1)(m-2)] d.f. at the level of significance α. Thus if $F_R > F_\alpha$ we reject $H_{0r}$ and if $F_R \leq F_\alpha$ we fail to reject $H_{0r}$. Similarly, we can test for $H_{0c}$ and $H_{0t}$.