

Chapter 14

Experimental Designs and Their Analysis

Design of experiment means how to design an experiment in the sense of how the observations or measurements should be obtained to answer a query in a valid, efficient, and economical way. The design of the experiment and the analysis of the obtained data are inseparable. If the experiment is properly designed, with the question in mind, then the data generated are valid, and proper analysis yields 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 terminology used in the experimental design.

Experimental unit

To conduct an experiment, the experimental material is divided into smaller parts, each of which is called an experimental unit. The experimental unit is randomly assigned to a treatment. The phrase “randomly assigned” is very important in this definition.

Experiment

A way of getting an answer to a question that the experimenter wants to know.

Treatment

Different objects or procedures that are to be compared in an experiment are called treatments.

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. A factor is termed a fixed factor if all the levels of interest are included in the experiment.

A factor is termed a 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.

Experimental error

The unexplained random part of the 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 arrangement of treatment levels in an experiment.

Example: (Ref.: *Statistical Design*, G. Casella, Chapman and Hall, 2008)

Suppose some varieties of fish food are to be investigated on some species of fish. The food is placed in water tanks containing fish. The response is the increase in fish weight. The experimental unit is the tank, as the treatment is applied to the tank rather than to the fish. Note that if the experimenter had taken the fish in hand and placed the food in the mouth of the fish, then the fish would have been the experimental unit as long as each of the fish got an independent scoop of food.

Design of experiment

One of the main objectives of designing an experiment is to verify the hypothesis efficiently and economically. In the context of the null hypothesis that several means of normal populations with equal variances are equal, 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 a hypothesis test may be invalid, and the data analysis may be meaningless. So, the main question is how to obtain data that meet the assumptions and are readily available for the application of tools like analysis of variance. The design of such a mechanism to obtain such data is achieved through the experimental design. After obtaining a sufficient number of experimental units, the treatments are randomly assigned to the units. Design of experiments provides a method for assigning treatments at random to experimental units so that responses are estimated with the utmost precision possible.

Principles of experimental design

There are three basic principles of design that were developed by Sir Ronald A. Fisher.

- (i) Randomization
- (ii) Replication
- (iii) Local control

(i) Randomization

The principle of randomization involves the allocation of treatment to experimental units at random to avoid any bias in the experiment resulting from the influence of some extraneous unknown factor that may affect the experiment. In the development of analysis of variance, we assume that the errors are random and independent. In turn, observations also become random. The principle of randomization ensures this.

The random assignment of experimental units to treatments results in the following outcomes.

- a) It eliminates systematic bias.
- b) It is needed to obtain a representative sample from the population.
- c) It helps in distributing the unknown variation due to confounded variables throughout the experiment and breaks the confounding influence.

Randomization is the basis of a valid experiment, but replication is also necessary for its validity.

If the randomization process is such that every experimental unit has an equal chance of receiving each treatment, it is called **complete randomization**.

(ii) Replication

In the replication principle, any treatment is repeated multiple times to obtain a more valid and reliable estimate than would be possible with a single observation. Replication provides an efficient way of increasing the precision of an experiment. Precision increases with the number of observations. Replication provides more observations when the same treatment is used, thereby increasing precision. For example, if

the variance of x is σ^2 then the variance of the sample mean $\bar{x}$ based on n observation is $\frac{\sigma^2}{n}$. So as n increases, $Var(\bar{x})$ decreases.

(iii) Local control (error control)

Replication is used with local control to reduce experimental errors. For example, if the experimental units are divided into groups such that they are homogeneous within blocks, then the variation among blocks is eliminated, and ideally, the error component will contain only the variation due to the treatments. This will, in turn, increase efficiency.

Complete and incomplete block designs

In most experiments, the available experimental units are grouped into blocks with more or less identical characteristics to remove the blocking effect from the experimental error. Such a design is termed a block design.

The number of experimental units in a block is called the **block size**.

If the size of the block is equal to the number of treatments and each treatment in each block is randomly allocated, then it is a **full replication**, and the design is called a **complete block design**.

If the number of treatments is so large that a full replication in each block would make the blocks too heterogeneous with respect to the characteristic under study, smaller, homogeneous blocks can be used. In such a case, the blocks do not contain a full replicate of the treatments. Experimental designs with blocks containing an incomplete replication of the treatments are called **incomplete block designs**.

Completely randomized design (CRD)

The CRD is the simplest design. Suppose there are v treatments to be compared.

- All experimental units are considered the same, and no division or grouping among them exists.
- In CRD, the v treatments are allocated randomly to the whole set of experimental units, without making any effort to group the experimental units in any way for more homogeneity.
- Design is entirely flexible in the sense that any number of treatments or replications may be used.
- The number of replications for different treatments need not be equal and may vary from treatment to treatment, depending on the knowledge (if any) on the variability of the observations on individual treatments, as well as on the accuracy required for the estimate of individual treatment effect.

Example: Suppose there are 4 treatments and 20 experimental units, then

- the treatment 1 is replicated, say 3 times and is given to 3 experimental units,
 - the treatment 2 is replicated, say 5 times and is given to 5 experimental units,
 - the treatment 3 is replicated, say 6 times and is given to 6 experimental units
 - and finally, the treatment 4 is replicated $[20-(6+5+3)]=6$ times and is given to the remaining 6 experimental units.
- All the variability among the experimental units goes into experimental error.
 - CRD is used when the experimental material is homogeneous.
 - CRD is often inefficient.
 - CRD is more useful when the experiments are conducted inside the lab.
 - CRD is well-suited for the small number of treatments and for the homogeneous experimental material.

Layout of CRD

The following steps are needed to design a CRD:

- Divide the entire experimental material or area into a number of experimental units, say n .
- Fix the number of replications for different treatments in advance (given a fixed total number of available experimental units).
- No local control measure is provided as such, except that the error variance can be reduced by choosing a homogeneous set of experimental units.

Procedure

Let the v treatments be numbered from 1,2,..., v and n_i be the number of replications required for i^{th}

treatment such that $\sum_{i=1}^v n_i = n$.

- Select n_1 units out of n units randomly and apply treatment 1 to these n_1 units.
(**Note:** This is how the randomization principle is utilised in CRD.)
- Select n_2 units out of $(n - n_1)$ units randomly and apply treatment 2 to these n_2 units.
- Continue with this procedure until all the treatments have been utilized.
- Generally, an equal number of treatments is allocated to all the experimental units unless there is no practical limitation or some treatments are more variable or/and/or of more interest.

Analysis

There is only one factor affecting the outcome – the treatment effect. So the setup of one-way analysis of variance is to be used.

y_{ij} : Individual measurement of j^{th} experimental units for i^{th} treatment $i = 1, 2, \dots, v, j = 1, 2, \dots, n_i$.

y_{ij} : Independently distributed following $N(\mu + \alpha_i, \sigma^2)$ with $\sum_{i=1}^v n_i \alpha_i = 0$.

μ : overall mean

α_i : i^{th} treatment effect

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_v = 0$$

$$H_1 : \text{All } \alpha_i \text{'s are not equal.}$$

The data set is arranged as follows:

Treatments				
1	2	...	v	
y_{11}	y_{21}	...	y_{v1}	
y_{12}	y_{22}	...	y_{v2}	
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
y_{1n_1}	y_{2n_2}	...	y_{vn_v}	
T_1	T_2	...	T_v	

where $T_i = \sum_{j=1}^{n_i} y_{ij}$ is the treatment total due to i^{th} effect, and $G = \sum_{i=1}^v T_i = \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij}$ is the grand total of all the observations.

In order to derive the test for H_0 , we can use either the likelihood ratio test or the principle of least squares.

The linear model under consideration is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, v, j = 1, 2, \dots, n_i$$

where ε_{ij} 's are identically and independently distributed random errors with mean 0 and variance σ^2 . The normality assumption of ε 's is not needed for the estimation of parameters but will be needed for deriving the distribution of various involved statistics and in deriving the test statistics.

The observations on Y and the design matrix X in the form of the model $Y = X\beta + \varepsilon$ in this case are arranged as

$$\begin{array}{c|ccc} Y & \mu & \alpha_1 & \alpha_2 & \cdots & \alpha_I \\ \hline y_{11} & 1 & 1 & 0 & \cdots & 0 \\ y_{12} & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ y_{1n_1} & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ y_{v1} & 1 & 0 & 0 & \cdots & 1 \\ y_{v2} & 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ y_{vn_j} & 1 & 0 & 0 & & 1 \end{array}$$

Let
$$S = \sum_{i=1}^v \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2.$$

Minimizing S with respect to μ and α_i , the normal equations are obtained as

$$\begin{aligned} \frac{\partial S}{\partial \mu} = 0 &\Rightarrow n\mu + \sum_{i=1}^v n_i \alpha_i = 0 \\ \frac{\partial S}{\partial \alpha_i} = 0 &\Rightarrow n_i \mu + n_i \alpha_i = \sum_{j=1}^{n_i} y_{ij} \quad i = 1, 2, \dots, v. \end{aligned}$$

Solving them using $\sum_{i=1}^v n_i \alpha_i = 0$, we get

$$\begin{aligned} \hat{\mu} &= \bar{y}_{oo} \\ \hat{\alpha}_i &= \bar{y}_{io} - \bar{y}_{oo} \end{aligned}$$

where $\bar{y}_{io} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{y}_{ij}$ is the mean of observations receiving the i^{th} treatment and $\bar{y}_{oo} = \frac{1}{n} \sum_{i=1}^v \sum_{j=1}^{n_i} \bar{y}_{ij}$ is the mean of all the observations.

The fitted model is obtained after substituting the estimate $\hat{\mu}$ and $\hat{\alpha}_i$ in the linear model. Using the fitted model, we can write

$$\begin{aligned} y_{ij} &= \bar{y}_{oo} + (\bar{y}_{io} - \bar{y}_{oo}) + (y_{ij} - \bar{y}_{io}) \\ \text{or } (y_{ij} - \bar{y}_{oo}) &= (\bar{y}_{io} - \bar{y}_{oo}) + (y_{ij} - \bar{y}_{io}). \end{aligned}$$

Squaring both sides and summing over all the observations, we have

$$\sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{oo})^2 = \sum_{i=1}^v n_i (\bar{y}_{io} - \bar{y}_{oo})^2 + \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{io})^2$$

$\downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow$

or $\left(\begin{array}{c} \text{Total sum} \\ \text{of squares} \end{array} \right) = \left(\begin{array}{c} \text{Sum of squares} \\ \text{due to treatment} \\ \text{effects} \end{array} \right) + \left(\begin{array}{c} \text{Sum of squares} \\ \text{due to error} \end{array} \right)$

or $TSS = SSTr + SSE$

- Since $\sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{oo}) = 0$, so TSS is based on the sum of $(n-1)$ squared quantities. The TSS carries only $(n-1)$ degrees of freedom.
- Since $\sum_{i=1}^v n_i (\bar{y}_{io} - \bar{y}_{oo}) = 0$, so $SSTr$ is based only on the sum of $(v-1)$ squared quantities. The $SSTr$ carries only $(v-1)$ degrees of freedom.
- Since $\sum_{i=1}^v n_i (\bar{y}_{ij} - \bar{y}_{io}) = 0$ for all $i = 1, 2, \dots, v$, so SSE is based on the sum of squaring n quantities like $(y_{ij} - \bar{y}_{io})$ with v constraints $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{io}) = 0$, So SSE carries $(n-v)$ degrees of freedom.
- Using the Fisher-Cochran theorem,
 $TSS = SSTr + SSE$
with degrees of freedom partitioned as
 $(n-1) = (v-1) + (n-v)$.

Moreover, equality in $TSS = SSTr + SSE$ has to hold exactly. To ensure that the equality holds exactly, we find one of the sums of squares through subtraction. Generally, it is recommended to find SSE by subtraction as

$$\begin{aligned} SSE &= TSS - SSTr \\ TSS &= \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{io})^2 \\ &= \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{n} \end{aligned}$$

where

$$G = \sum_{i=1}^v \sum_{j=1}^{n_i} y_{ij}.$$

$$\begin{aligned} SSTr &= \sum_{j=1}^{n_i} n_i (\bar{y}_{io} - \bar{y}_{oo})^2 \\ &= \sum_{i=1}^v \left(\frac{T_i^2}{n_i} \right) - \frac{G^2}{n} \end{aligned}$$

where $T_i = \sum_{j=1}^{n_i} y_{ij}$ and $\frac{G^2}{n}$ is the correction factor..

Now under $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_v = 0$, the model becomes

$$Y_{ij} = \mu + \varepsilon_{ij},$$

and minimizing $S = \sum_{i=1}^v \sum_{j=1}^{n_i} \varepsilon_{ij}^2$ with respect to μ gives

$$\frac{\partial S}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{G}{n} = \bar{y}_{oo}.$$

The SSE under H_0 becomes

$$SSE = \sum_{i=1}^v \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{oo})^2$$

and thus $TSS = SSE$. This TSS under H_0 contains the variation only due to the random error, whereas the earlier $TSS = SSTr + SSE$ contains the variation due to treatments and errors. The difference between the two will provide the effect of treatments in terms of the sum of squares as

$$SSTr = \sum_{i=1}^v n_i (\bar{y}_i - \bar{y}_{oo})^2$$

Distributions and decision rules

Using the normal distribution property of ε_{ij} 's, we find that y_{ij} 's are also normal as they are the linear combination of ε_{ij} 's.

$$-\frac{SSTr}{\sigma^2} \sim \chi^2(v-1) \text{ under } H_0$$

$$-\frac{SSE}{\sigma^2} \sim \chi^2(n-v) \text{ under } H_0$$

– $SSTr$ and SSE are independently distributed

– $\frac{MStr}{MSE} \sim F(v-1, n-v)$ under H_0 .

– Reject H_0 at α^* level of significance if $F > F_{\alpha^*; v-1, n-v}$.

[Note: We denote the level of significance here by α^* because α has been used for denoting the factor]

The analysis of variance table is as follows:

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F
Between treatments	$v - 1$	$SSTr$	$MStr$	$\frac{MStr}{MSE}$
Errors	$n - v$	SSE	MSE	
Total	$n - 1$	TSS		

Expectations

$$\begin{aligned}
 E(SSE) &= \sum_{i=1}^v \sum_{j=1}^{n_i} E(y_{ij} - y_{io})^2 \\
 &= \sum_{i=1}^v \sum_{j=1}^{n_i} (\varepsilon_{ij} - \bar{\varepsilon}_{io})^2 \\
 &= \sum_{i=1}^v \sum_{j=1}^{n_i} E(\varepsilon_{ij}^2) - \sum_{i=1}^v n_i E(\bar{\varepsilon}_{io}^2) \\
 &= n\sigma^2 - \sum_{i=1}^v n_i \frac{\sigma^2}{n_i} \\
 &= (n - v)\sigma^2
 \end{aligned}$$

$$E(MSE) = E\left(\frac{SSE}{n - v}\right) = \sigma^2$$

$$\begin{aligned}
E(SSTr) &= \sum_{i=1}^v n_i E(\bar{y}_{io} - \bar{y}_{oo})^2 \\
&= \sum_{i=1}^v n_i E(\alpha_i + \bar{\varepsilon}_{io} - \bar{\varepsilon}_{oo})^2 \\
&= \sum_{i=1}^v n_i \alpha_i^2 + \left[\sum_{i=1}^v n_i \bar{\varepsilon}_{io}^2 - n \bar{\varepsilon}_{oo}^2 \right] \\
&= \sum_{i=1}^v n_i \alpha_i^2 + \left[\sum_{i=1}^v n_i \frac{\sigma^2}{n_i} - n \frac{\sigma^2}{n} \right] \\
&= \sum_{i=1}^v n_i \alpha_i^2 + (v-1)\sigma^2
\end{aligned}$$

$$E(MSTr) = E\left(\frac{SSTr}{v-1}\right) = \frac{1}{v-1} \sum_{i=1}^v n_i \alpha_i^2 + \sigma^2.$$

In general, $E(MSTr) \neq \sigma^2$ but under H_0 , all $\alpha_i = 0$ and so $E(MSTr) = \sigma^2$.

Randomized Block Design

If a large number of treatments are to be compared, then a large number of experimental units are required. This will increase the variation among the responses, and CRD may not be appropriate to use. In such a case, when the experimental material is not homogeneous and there are v treatments to be compared, then it may be possible to

- group the experimental material into blocks of sizes v units.
- Blocks are constructed such that the experimental units within a block are relatively homogeneous and resemble each other more closely than the units in the different blocks.
- If there are b such blocks, we say that the blocks are at b levels. Similarly, if there are v treatments, we say that the treatments are at v levels. The responses from the b levels of blocks and v levels of treatments can be arranged in a two-way layout. The observed data set is arranged as follows:

	Treatments (Factor B)							Block totals
		1	2		j		v	
Blocks (Factor A)	1	y_{11}	y_{12}	...	y_{1j}	...	y_{1v}	B_1
	2	y_{21}	y_{22}	...	y_{2j}	...	y_{2v}	B_2
	.	.	.		.		.	.
	.	.	.		.		.	.
	.	.	.		.		.	.
	i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{iv}	B_i
	.		.		.		.	.
	.		.		.		.	.
	.		.		.		.	.
	b	y_{b1}	y_{b2}	...	y_{bj}	...	y_{bv}	B_b
Treatment totals		T_1	T_2	...	T_j	...	T_v	Grand total (G)

Layout

A two-way layout is called a randomized block design (RBD) or a randomized complete block design (RCB) if, within each block, the v treatments are randomly assigned to v experimental units such that each of the $v!$ ways of assigning the treatments to the units has the same probability of being adopted in the experiment and the assignments in different blocks are statistically independent.

The RBD utilizes the principles of design - randomization, replication and local control - in the following way:

1. Randomization

- Number the v treatments $1, 2, \dots, v$.
- Number the units in each block as $1, 2, \dots, v$.
- Randomly allocate the v treatments to v experimental units in each block.

2. Replication

Since each treatment appears in each block, every treatment will appear in all the blocks. So each treatment can be considered as if it were replicated the number of times as the number of blocks. Thus, in RBD, the number of blocks and the number of replications are the same.

3. Local control

Local control is adopted in RBD in the following way:

- First form the homogeneous blocks of the experimental units.
- Then allocate each treatment randomly in each block.

The error variance now will be smaller because of homogeneous blocks and some variance will be parted away from the error variance due to the difference among the blocks.

Example

Suppose there are 7 treatments denoted as $T_1, T_2, \dots, T_7$ corresponding to 7 levels of a factor to be included in 4 blocks. So, one possible layout of the assignment of 7 treatments to 4 different blocks in an RBD is as follows:

Block 1	T_2	T_7	T_3	T_5	T_1	T_4	T_6
Block 2	T_1	T_6	T_7	T_4	T_5	T_3	T_2
Block 3	T_7	T_5	T_1	T_6	T_4	T_2	T_3
Block 4	T_4	T_1	T_5	T_6	T_2	T_7	T_3

Analysis

Let

y_{ij} : Individual measurements of j^{th} treatment in i^{th} block, $i = 1, 2, \dots, b$, $j = 1, 2, \dots, v$.

y_{ij} 's are independently distributed following $N(\mu + \beta_i + \tau_j, \sigma^2)$

where μ : overall mean effect

β_i : i^{th} block effect

τ_j : j^{th} treatment effect

such that $\sum_{i=1}^b \beta_i = 0$, $\sum_{j=1}^v \tau_j = 0$.

There are two null hypotheses to be tested.

- related to the block effects $H_{0B} : \beta_1 = \beta_2 = \dots = \beta_b = 0$.
- related to the treatment effects $H_{0T} : \tau_1 = \tau_2 = \dots = \tau_v = 0$.

The linear model, in this case, is a two-way model as

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, i = 1, 2, \dots, b; j = 1, 2, \dots, v$$

where ε_{ij} are identically and independently distributed random errors following a normal distribution with mean 0 and variance σ^2 .

The observations on Y and the design matrix X in the form of the model $Y = X\beta + \varepsilon$ in this case are arranged as

$$\begin{array}{c|cccc|cccc} Y & \mu & \beta_1 & \beta_2 & \cdots & \beta_b & \tau_1 & \tau_2 & \cdots & \tau_v \\ \hline y_{11} & 1 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ y_{12} & 1 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & & & \\ y_{1v} & 1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ y_{b1} & 1 & 0 & 0 & \cdots & 1 & 1 & 0 & & 0 \\ y_{b2} & 1 & 0 & 0 & & 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & & \vdots \\ y_{bv} & 1 & 0 & 0 & & 1 & 0 & 0 & & 1 \end{array}$$

The tests of hypotheses can be derived using the likelihood ratio test or the principle of least squares. The use of the likelihood ratio test has already been demonstrated earlier, so we now use the principle of least squares.

$$\text{Minimizing } S = \sum_{i=1}^b \sum_{j=1}^v \varepsilon_{ij}^2 = \sum_{i=1}^b \sum_{j=1}^v (y_{ij} - \mu - \beta_i - \tau_j)^2$$

and solving the normal equation

$$\frac{\partial S}{\partial \mu} = 0, \frac{\partial S}{\partial \beta_i} = 0, \frac{\partial S}{\partial \tau_j} = 0 \text{ for all } i = 1, 2, \dots, b, j = 1, 2, \dots, v,$$

the least squares estimators are obtained as

$$\begin{aligned} \hat{\mu} &= \bar{y}_{oo}, \\ \hat{\beta}_i &= \bar{y}_{io} - \bar{y}_{oo}, \\ \hat{\tau}_j &= \bar{y}_{oj} - \bar{y}_{oo}. \end{aligned}$$

Using the fitted model (obtained after substituting the estimated values of the parameters in the model), we can write

$$y_{ij} = \bar{y}_{oo} + (\bar{y}_{io} - \bar{y}_{oo}) + (\bar{y}_{oj} - \bar{y}_{oo}) + (y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo}).$$

Squaring both sides and summing over i and j gives

$$\sum_{i=1}^b \sum_{j=1}^v (\bar{y}_{ij} - \bar{y}_{oo})^2 = v \sum_{i=1}^b (\bar{y}_{io} - \bar{y}_{oo})^2 + b \sum_{j=1}^v (\bar{y}_{oj} - \bar{y}_{oo})^2 + \sum_{i=1}^b \sum_{j=1}^v (y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo})^2$$

or $TSS = SSBl + SSTr + SSE$

with degrees of freedom partitioned as

$$bv - 1 = (b - 1) + (v - 1) + (b - 1)(v - 1).$$

The reason for the number of degrees of freedom for different sums of squares is the same as in the case of CRD.

Here

$$\begin{aligned} TSS &= \sum_{i=1}^b \sum_{j=1}^v (y_{ij} - \bar{y}_{oo})^2 \\ &= \sum_{i=1}^b \sum_{j=1}^v y_{ij}^2 - \frac{G^2}{bv} \end{aligned}$$

where $\frac{G^2}{bv}$ is the correction factor and $G = \sum_{i=1}^b \sum_{j=1}^v y_{ij}$ is the grand total of all the observations.

$$\begin{aligned} SSBl &= v \sum_{i=1}^b (\bar{y}_{io} - \bar{y}_{oo})^2 \\ &= \sum_{i=1}^b \frac{B_i^2}{v} - \frac{G^2}{bv} \end{aligned}$$

where $B_i = \sum_{j=1}^v y_{ij}$ is the i^{th} block total,

$$\begin{aligned} SSTr &= b \sum_{j=1}^v (\bar{y}_{oj} - \bar{y}_{oo})^2 \\ &= \sum_{j=1}^v \frac{T_j^2}{b} - \frac{G^2}{bv} \end{aligned}$$

where $T_j = \sum_{i=1}^b y_{ij}$ is the j^{th} treatment total.

$$SSE = \sum_{i=1}^b \sum_{j=1}^v (y_{ij} - \bar{y}_{io} - \bar{y}_{oj} + \bar{y}_{oo})^2.$$

The expectations of the mean squares are

$$E(MSBl) = E\left(\frac{SSBl}{b-1}\right) = \sigma^2 + \frac{v}{b-1} \sum_{i=1}^b \beta_i^2$$

$$E(MSTr) = E\left(\frac{SSTr}{v-1}\right) = \sigma^2 + \frac{b}{v-1} \sum_{j=1}^v \tau_j^2$$

$$E(MSE) = E\left(\frac{SSE}{(b-1)(v-1)}\right) = \sigma^2.$$

Moreover,

$$(b-1) \frac{SSBl}{\sigma^2} \sim \chi^2(b-1)$$

$$(v-1) \frac{SSTr}{\sigma^2} \sim \chi^2(v-1)$$

$$(b-1)(v-1) \frac{SSE}{\sigma^2} \sim \chi^2(b-1)(v-1).$$

Under $H_{0B} : \beta_1 = \beta_2 = \dots = \beta_b = 0$,

$$E(MSBl) = E(MSE)$$

and $SSBl$ and SSE are independent, so

$$F_{bl} = \frac{MSBl}{MSE} \sim F((b-1), (b-1)(v-1)).$$

Similarly, under $H_{0T} : \tau_1 = \tau_2 = \dots = \tau_v = 0$, so

$$E(MSTr) = E(MSE)$$

and $SSTr$ and SSE are independent, so

$$F_{Tr} = \frac{MSTr}{MSE} \sim F(v-1, (b-1)(v-1)).$$

The decision rules are:

Reject H_{0B} if $F_{be} > F_{\alpha}((b-1), (b-1)(v-1))$

Reject H_{0T} if $F_{Tr} > F_{\alpha}((v-1), (b-1)(v-1)).$

If H_{0B} is accepted, then it indicates that the blocking is not necessary for future experimentation.

If H_{0T} is rejected, then it indicates that the treatments are different. Then the multiple comparison tests are used to divide the entire set of treatments into different subgroups such that the treatments in the same subgroup have the same treatment effect and those in the different subgroups have different treatment effects.

The analysis of variance table is as follows

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Blocks	$b - 1$	SS_{Bl}	MS_{Bl}	F_{Bl}
Treatments	$v - 1$	SS_{Tr}	MS_{Tr}	F_{Tr}
Errors	$(b - 1)(v - 1)$	SSE	MSE	
Total	$bv - 1$	TSS		

Latin Square Design

The treatments in the RBD are randomly assigned to b blocks, such that each treatment occurs in each block, rather than being assigned at random across the entire set of experimental units as in the CRD. There are only two factors – block and treatment effects – which are taken into account and the total number of experimental units needed for complete replication are bv , where b and v are the numbers of blocks and treatments respectively.

If there are three factors and there are b , v , and k levels of each factor, then the total number of experimental units needed for a complete replication is bvk . This increases the cost of experimentation and the required number of experimental units over RBD.

In Latin square design (LSD), the experimental material is divided into rows and columns, each having the same number of experimental units, which is equal to the number of treatments. The treatments are allocated to the rows and columns such that each treatment occurs exactly once in each row and exactly once in each column.

To allocate the treatment to the experimental units in rows and columns, we use Latin squares.

Latin Square

A Latin square of order p is an arrangement of p symbols in p^2 cells arranged in p rows and p columns such that each symbol occurs once and only once in each row and in each column. For example, to write a Latin square of order 4, choose four symbols – A, B, C and D. These letters are Latin letters which are used as symbols. Write them in a way such that each of the letters out of A, B, C and D occurs once and only once in each row and each column. For example, as

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

This is a Latin square.

We first consider the following example to illustrate how a Latin square is used to allocate treatments and obtain the response.

Example

Suppose different brands of petrol are to be compared with respect to the mileage per litre achieved in motor cars.

Important factors responsible for the variation in mileage are

- the difference between individual cars.
- the difference in the driving habits of drivers.

We have three factors – cars, drivers and petrol brands. Suppose we have

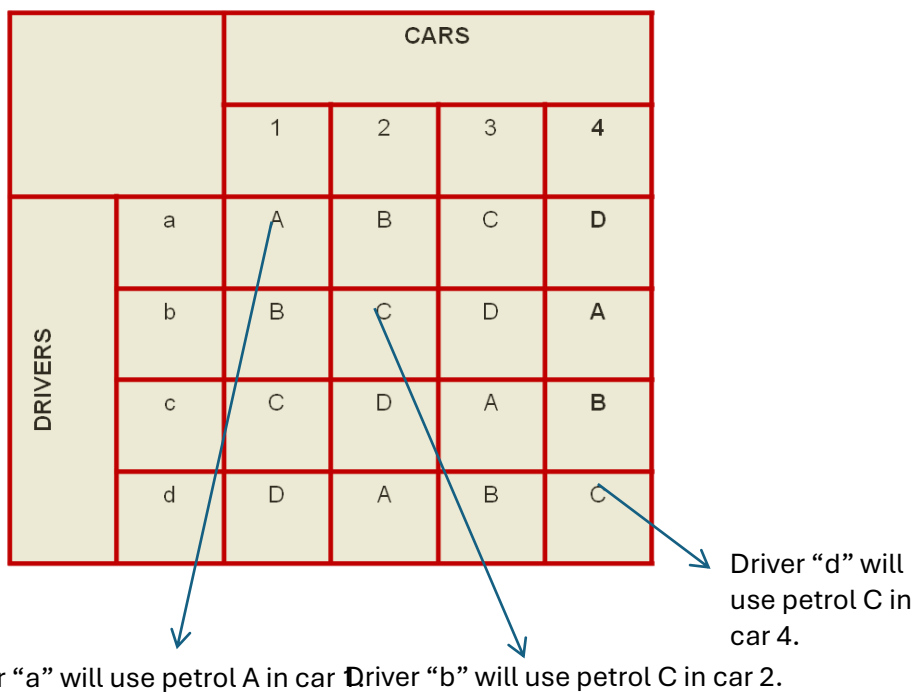
- 4 types of cars denoted as 1, 2, 3, 4.
- 4 drivers that are represented by a, b, c, d.
- 4 brands of petrol are indicated as A, B, C, D.

Now the complete replication will require $4 \times 4 \times 4 = 64$ the number of experiments. We chose only 16 experiments.

To select these 16 experiments, we use a Latin square. Suppose we choose the following Latin square:

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Write them in rows and columns and choose rows for drivers, columns for cars and letters for petrol brands. Thus, 16 observations are recorded according to this treatment combination plan (as shown in the next figure), and further analysis is carried out. Since such a design is based on a Latin square, it is called a Latin square design.



Another choice of a Latin square of order 4 is

C	B	A	D
B	C	D	A
A	D	C	B
D	A	B	C

This will again give a design different from the previous one. The 16 observations will be recorded again, but using different treatment combinations.

Since we use only 16 out of 64 possible observations, it is an incomplete 3-way layout in which each of the 3 factors – cars, drivers, and petrol brands - is at 4 levels, and observations are recorded only for 16 of the 64 possible treatment combinations.

Thus, in an LSD,

- the treatments are grouped into replications in two-ways
 - once in rows and
 - and in columns,
- rows and columns variations are eliminated from the within treatment variation.
 - In RBD, the experimental units are divided into homogeneous blocks according to the blocking factor. Hence, it eliminates differences among blocks due to experimental error.
 - In LSD, the experimental units are grouped according to two factors. Hence, two effects (like two block effects) are removed from the experimental error.
 - So the error variance can be considerably reduced in LSD.

The LSD is an incomplete three-way layout in which each of the three factors, viz., rows, columns, and treatments, is at v levels each, and observations are only on v^2 of the v^3 possible treatment combinations are taken. Each treatment combination contains one level of each factor.

The analysis of data in an LSD is conditional in the sense that it depends on which Latin square is used for allocating the treatments. If the Latin square changes, the conclusions may also change.

We note that Latin squares play an important role in an LSD, so first we study more about these Latin squares before describing the analysis of variance.

Standard form of Latin square

A Latin square is in the standard form if the symbols in the first row and first column are in the natural order (Natural order means the order of the alphabet, like A, B, C, D,...).

Given a Latin square, it is possible to rearrange the columns so that the first row and first column remain in a natural order.

Example: Four standard forms of 4×4 Latin squares are as follows.

A B C D	A B C D	A B C D	A B C D
B A D C	B C D A	B D A C	B A D C
C D B A	C D A B	C A D B	C D A B
D C A B	D A B C	D C B A	D C B A

For each standard Latin square of order p , the p rows can be permuted in $p!$ ways. Keeping a row fixed, vary and permute $(p - 1)$ columns in $(p - 1)!$ ways. So there are $p!(p - 1)!$ different Latin squares.

For illustration

Size the the square	Number of Standard squares	Value of $p!(1 - p)!$	Total number of different squares
3 $\times$ 3	1	12	12
4 $\times$ 4	4	144	576
5 $\times$ 5	56	2880	161280
6 $\times$ 6	9408	86400	812851250

Conjugate

Two standard Latin squares are called conjugate if the rows of one are the columns of other.

For example,

A B C D		A B C D
B C D A	and	B C D A
C D A B		C D A B
D A B C		D A B C

are conjugate. In fact, they are **self conjugate**.

A Latin square is called **self conjugate** if its arrangement in rows and columns is the same.

Transformation set

A set of all Latin squares obtained from a single Latin square by permuting its rows, columns and symbols is called a transformation set.

From a Latin square of order p , $p!(p - 1)!$ different Latin squares can be obtained by making $p!$ permutations of columns and $(p - 1)!$ permutations of rows which leaves the first row in place. Thus

Number of different Latin squares of order p in a transformation set = $p!(p - 1)!$ X number of standard Latin squares in the set

Orthogonal Latin squares

If two Latin squares of the same order but with different symbols are such that when they are superimposed on each other, every ordered pair of symbols (different) occurs exactly once in the Latin square, then they are called orthogonal.

Graeco-Latin square

A pair of orthogonal Latin squares, one with Latin symbols and the other with Greek symbols, forms a Graeco-Latin square.

For example

A	B	C	D	α	β	γ	δ
B	A	D	C	δ	γ	β	α
C	D	A	B	β	α	δ	γ
D	C	B	A	γ	δ	α	β

is a Graeco-Latin square of order 4.

The Graeco-Latin squares design enables the consideration of one more factor than the factors in the Latin square design. For example, in the earlier example, if there are four drivers, four cars, four types of petrol, and each petrol has four varieties, such as α, β, γ and δ , then the Graeco-Latin square helps in deciding the treatment combination as follows:

Drivers	Cars				
		1	2	3	4
	<i>a</i>	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
	<i>b</i>	$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
	<i>c</i>	$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
	<i>d</i>	$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

Now

$A\alpha$ means: Driver 'a' will use the α variant of petrol A in Car 1.

$B\gamma$ means: Driver 'c' will use the γ variant of petrol B in Car 4

and so on.

Mutually orthogonal Latin square

A set of Latin squares of the same order is called a set of mutually orthogonal Latin squares (or a hyper Graeco-Latin square) if every pair in the set is orthogonal. The total number of mutually orthogonal Latin squares of order p is at most $(p - 1)$.

Analysis of LSD (one observation per cell)

In designing an LSD of order p ,

- choose one Latin square at random from the set of all possible Latin squares of order p .
- Select a standard Latin square from the set of all standard Latin squares with equal probability.
- Randomize all the rows and columns as follows:
 - Choose a random number, less than p , say n_1 and then 2nd row is the n_1^{th} row.
 - Choose another random number less than p , say n_2 and then 3rd row is the n_2^{th} row and so on.
 - Then do the same for the column.
- For Latin squares of the order less than 5, fix the first row and then randomize rows and then randomize columns. In Latin squares of order 5 or more, it need not fix even the first row. Just randomize all rows and columns.

Example

Suppose the following Latin square is chosen

A B C D E
B C D E A
D E A B C
E A B C D
C D E A B

Now randomize rows, e.g., 3rd row becomes 5th row and 5th row becomes 3rd row. The Latin square becomes

A B C D E
B C D E A
C D E A B
E A B C D
D E A B C.

Now randomize columns, say 5th column becomes 1st column, 1st column becomes 4th column and 4th column becomes 5th column

E B C A D
A C D B E
D A B E C
C E A D B
B D E C A

Now use this Latin square for the assignment of treatments.

y_{ijk} : Observation on k^{th} treatment in i^{th} row and j^{th} block, $i = 1, 2, \dots, v$, $j = 1, 2, \dots, v$, $k = 1, 2, \dots, v$.

Triplets (i, j, k) take on only the v^2 values indicated by the chosen Latin square selected for the experiment.

y_{ijk} 's are independently distributed as $N(\mu + \alpha_i + \beta_j + \tau_k, \sigma^2)$.

A linear model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{ijk}, i = 1, 2, \dots, v; j = 1, 2, \dots, v; k = 1, 2, \dots, v$$

where ε_{ijk} are random errors which are identically and independently distributed following $N(0, \sigma^2)$.

with $\sum_{i=1}^v \alpha_i = 0$, $\sum_{j=1}^v \beta_j = 0$, $\sum_{k=1}^v \tau_k = 0$,

α_i : main effect of rows

β_j : main effect of columns

γ_k : main effect of treatments.

The null hypothesis under consideration are

$$H_{0R} : \alpha_1 = \alpha_2 = \dots = \alpha_v = 0$$

$$H_{0C} : \beta_1 = \beta_2 = \dots = \beta_v = 0$$

$$H_{0T} : \tau_1 = \tau_2 = \dots = \tau_v = 0$$

The analysis of variance can be developed on the same lines as earlier.

Minimizing $S = \sum_{i=1}^v \sum_{j=1}^v \sum_{k=1}^v \varepsilon_{ijk}^2$ with respect to μ, α_i, β_j and τ_k given the least-squares estimate as

$$\hat{\mu} = \bar{y}_{ooo}$$

$$\hat{\alpha}_i = \bar{y}_{ioo} - \bar{y}_{ooo} \quad i = 1, 2, \dots, v$$

$$\hat{\beta}_j = \bar{y}_{ojo} - \bar{y}_{ooo} \quad j = 1, 2, \dots, v$$

$$\hat{\tau}_k = \bar{y}_{ook} - \bar{y}_{ooo} \quad k = 1, 2, \dots, v.$$

Using the fitted model based on these estimators, the total sum of squares can be partitioned into the mutually orthogonal sum of squares SSR , SSC , $SSTr$ and SSE as

$$TSS = SSR + SSC + SSTr + SSE$$

where

$$TSS: \text{Total sum of squares} = \sum_{i=1}^v \sum_{j=1}^v \sum_{k=1}^v (y_{ijk} - \bar{y}_{ooo})^2 = \sum_{i=1}^v \sum_{j=1}^v \sum_{k=1}^v y_{ijk}^2 - \frac{G^2}{v^2}$$

$$SSR: \text{Sum of squares due to rows} = v \sum_{i=1}^v (\bar{y}_{ioo} - \bar{y}_{ooo})^2 = \frac{\sum_{i=1}^v R_i^2}{v} - \frac{G^2}{v^2}; \text{ where } R_i = \sum_{j=1}^v \sum_{k=1}^v y_{ijk}$$

$$SSC: \text{Sum of squares due to column} = v \sum_{j=1}^v (\bar{y}_{ojo} - \bar{y}_{ooo})^2 = \frac{\sum_{j=1}^v C_j^2}{v} - \frac{G^2}{v^2}; \text{ where } C_j = \sum_{i=1}^v \sum_{k=1}^v y_{ijk}$$

$$SSTr : \text{Sum of squares due to treatment} = v \sum_{k=1}^v (\bar{y}_{ook} - \bar{y}_{ooo})^2 = \frac{\sum_{k=1}^v T_k^2}{v} - \frac{G^2}{v^2}; \text{ where } T_k = \sum_{i=1}^v \sum_{j=1}^v y_{ijk}$$

Degrees of freedom carried by SSR , SSC and $SSTr$ are $(v - 1)$ each.

Degrees of freedom carried by TSS is $v^2 - 1$.

Degree of freedom carried by SSE is $(v - 1)(v - 2)$.

The expectations of the mean squares are obtained as

$$E(MSR) = E\left(\frac{SSR}{v-1}\right) = \sigma^2 + \frac{v}{v-1} \sum_{i=1}^v \alpha_i^2$$

$$E(MSC) = E\left(\frac{SSC}{v-1}\right) = \sigma^2 + \frac{v}{v-1} \sum_{j=1}^v \beta_j^2$$

$$E(MSTr) = E\left(\frac{SSTr}{v-1}\right) = \sigma^2 + \frac{v}{v-1} \sum_{k=1}^v \tau_k^2$$

$$E(MSE) = E\left(\frac{SSE}{(v-1)(v-2)}\right) = \sigma^2.$$

Thus

$$\text{- under } H_{0R}, F_R = \frac{MSR}{MSE} \sim F((v-1), (v-1)(v-2))$$

$$\text{- under } H_{0C}, F_C = \frac{MSC}{MSE} \sim F((v-1), (v-1)(v-2))$$

$$\text{- under } H_{0T}, F_T = \frac{MSTr}{MSE} \sim F((v-1), (v-1)(v-2)).$$

Decision rules

Reject H_{0R} at level α if $F_R > F_{1-\alpha; v(-1), (v-1)(v-2)}$

Reject H_{0C} at level α if $F_C > F_{1-\alpha; (v-1), (v-1)(v-2)}$

Reject H_{0T} at level α if $F_T > F_{1-\alpha; (v-1), (v-1)(v-2)}.$

If any null hypothesis is rejected, then use the multiple comparison test.

The analysis of variance table is as follows:

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F
Rows	$v - 1$	SSR	MSR	F_R
Columns	$v - 1$	SSC	MSC	F_C
Treatments	$v - 1$	$SSTr$	$MSTr$	F_T
Error	$(v - 1)(v - 2)$	SSE	MSE	
Total	$v^2 - 1$	TSS		

Missing plot techniques

In conducting experiments, it is common for some observations to be missed. This may happen due to several reasons. For example, in a clinical trial, suppose blood pressure readings are to be recorded after three days of administering the medicine to the patients. Suppose the medicine is given to 20 patients, and one of the patients doesn't turn up for providing the blood pressure reading. Similarly, in an agricultural experiment, seeds are sown, and yields are recorded after a few months. Suppose some cattle destroy the crop of any plot, or the crop of any plot is destroyed due to a storm, insects, etc.

In such cases, one option is to

- somehow estimate the missing value on the basis of available data,
- replace it back in the data and make the data set complete.

Now conduct the statistical analysis based on the completed data set, as if no values were missing, by making the necessary adjustments to the statistical tools to be applied. Such an area falls under the purview of “missing data models”, and considerable development has taken place. Several books on this issue have appeared, e.g.,

- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London etc.

We discuss here the classical missing plot technique proposed by Yates, which involves the following steps:

- Estimate the missing observations by the values that make the error sum of squares a minimum.
- Substitute the unknown values with the missing observations.
- Express the error sum of squares as a function of these unknown values.
- Minimize the error sum of squares using the principle of maxima/minima, i.e., differentiating it with respect to the missing value and setting it to zero and form a linear equation.
- Form as many linear equations as the number of unknown values (i.e., differentiate the error sum of squares with respect to each unknown value).
- Solve all the linear equations simultaneously, and the solutions will provide the missing values.
- Impute the missing values with the estimated values and complete the data.
- Apply analysis of variance tools.
- The error sum of squares thus obtained is corrected, but the treatment sum of squares is not corrected.

- The number of degrees of freedom associated with the total sum of squares is subtracted by the number of missing values and adjusted in the error sum of squares. No change in the degrees of freedom of the sum of squares due to treatment is needed.

Missing observations in RBD

One missing observation

Suppose one observation in the $(i, j)^{\text{th}}$ cell is missing and let this be x . The arrangement of observations in RBD will then look as follows:

	Treatments (Factor B)							Block totals
		1	2		j		v	
Blocks (Factor A)	1	y_{11}	y_{12}	...	y_{1j}	...	y_{1v}	$B_1 = y_{1o}$
	2	y_{21}	y_{22}	...	y_{2j}	...	y_{2v}	$B_2 = y_{2o}$
	.	.	.		.		.	.
	.	.	.		.		.	.
	.	.	.		.		.	.
	i	y_{i1}	y_{i2}	...	$y_{ij} = x$	...	y_{iv}	$B_i = y'_{io} + x$
	.		.		.		.	.
	.		.		.		.	.
	.		.		.		.	.
	b	y_{b1}	y_{b2}	...	y_{bj}	...	y_{bv}	B_b
Treatment totals		$T_1 = y_{o1}$	$T_2 = y_{o2}$	...	$T_j = y'_{oj} + x$	...	$T_v = y_{ob}$	Grand total (G) = $y'_{oo} + x$

where y'_{oo} : total of known observations

y'_{io} : total of known observations in i^{th} block

y'_{oj} : total of known observations in j^{th} treatment

$$\text{Correction factor (CF)} = \frac{(G')^2}{n} = \frac{(y'_{oo} + x)^2}{bv}$$

$$TSS = \sum_{i=1}^b \sum_{j=1}^v y_{ij}^2 - CF$$

$$= (x^2 + \text{terms which are constant with respect to } x) - CF$$

$$SSBl = \frac{1}{v}[(y'_{io} + x)^2 + \text{terms which are constant with respect to } x] - CF$$

$$SSTr = \frac{1}{b}[(y'_{oj} + x)^2 + \text{terms which are constant with respect to } x] - CF$$

$$SSE = TSS - SSBl - SSTr$$

$$= x^2 - \frac{1}{v}(y'_{io} + x)^2 - \frac{1}{b}(y'_{oj} + x)^2 + \frac{(y'_{oo} + x)^2}{bv} + (\text{terms which are constant with respect to } x) - CF.$$

Find x such that SSE is minimum

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

$$\Rightarrow 2x - \frac{2(y'_{io} + x)}{v} - \frac{2(y'_{oj} + x)}{b} + \frac{2(y'_{oo} + x)}{bv} = 0$$

$$\text{or } x = \frac{vy'_{oj} + by'_{io} - y'_{oo}}{(b-1)(v-1)}$$

The second-order derivative condition for x to provide a minimum SSE can be easily verified.

Two missing observations

If there are two missing observations, then let them be x and y .

- Let the corresponding row sums (block totals) be $(R_1 + x)$ and $(R_2 + y)$.
- Column sums (treatment totals) are $(C_1 + x)$ and $(C_2 + y)$.
- Total of known observations is S .

Then

$$SSE = x^2 + y^2 - \frac{1}{v}[(R_1 + x)^2 + (R_2 + y)^2] - \frac{1}{b}[(C_1 + x)^2 + (C_2 + y)^2] + \frac{1}{bv}(S + x + y)^2$$

+ terms independent of x and y .

Now differentiate SSE with respect to x and y , as

$$\frac{\partial(SSE)}{\partial x} = 0 \Rightarrow x - \frac{R_1 + x}{v} - \frac{C_1 + x}{b} + \frac{S + x + y}{bv} = 0$$

$$\frac{\partial(SSE)}{\partial y} = 0 \Rightarrow y - \frac{R_2 + y}{v} - \frac{C_2 + y}{b} + \frac{S + x + y}{bv} = 0.$$

Thus, solving the following two linear equations in x and y , we obtain the estimated missing values,

$$(b-1)(v-1)x = bR_1 + vC_1 - S - y$$

$$(b-1)(v-1)y = bR_2 + vC_2 - S - x.$$

Adjustments to be made in the analysis of variance

- (i) Obtain the within block sum of squares from incomplete data.
- (ii) Subtract the correct error sum of squares from (i). This gives the correct treatment sum of squares.
- (iii) Reduce the degrees of freedom of the error sum of squares by the number of missing observations.
- (iv) No adjustments in other sums of squares are required.

Missing observations in LSD

Let

- x be the missing observation in the $(i, j, k)^{\text{th}}$ cell, i.e., $y_{ijk}, i = 1, 2, \dots, v, j = 1, 2, \dots, v, k = 1, 2, \dots, v$.
- R : Total of known observations in i^{th} row
- C : Total of known observations in j^{th} column
- T : Total of known observations receiving the k^{th} treatment.
- S : Total of known observations

Now

$$\text{Correction factor (CF)} = \frac{(S + x)^2}{v^2}$$

Total sum of squares (TSS) = x^2 + term which are constant with respect to x - CF

$$\text{Row sum of squares (SSR)} = \frac{(R + x)^2}{v} + \text{term which are constant with respect to } x - CF$$

$$\text{Column sum of squares (SSC)} = \frac{(C + x)^2}{v} + \text{term which are constant with respect to } x - CF$$

$$\text{Treatment sum of squares (SSTr)} = \frac{(T + x)^2}{v} + \text{term which are constant with respect to } x - CF$$

$$\text{Sum of squares due to error (SSE)} = TSS - SSR - SSC - SSTr$$

$$= x^2 - \frac{1}{v} \left[(R + x)^2 + (C + x)^2 + (T + x)^2 \right] + \frac{2(S + x)^2}{v^2}$$

Choose x such that SSE is minimum. So

$$\frac{d(SSE)}{dx} = 0$$

$$\Rightarrow 2x - \frac{2}{v}(R + C + T + 3x) + \frac{4(S + x)}{v^2} = 0$$

$$\text{or } x = \frac{V(R + C + T) - 2S}{(v-1)(v-2)}$$

Adjustment to be done in the analysis of variance

Do all the steps as in the case of RBD.

To get the correct treatment sum of squares, proceed as follows:

- Ignore the treatment classification and consider only row and column classification.
- Substitute the estimated values for the missing observations.
- Obtain the error sum of squares from complete data, say SSE_1 .
- Let SSE_2 be the error sum of squares based on LSD obtained earlier.
- Find the corrected treatment sum of squares = $SSE_2 - SSE_1$.
- Reduce the degrees of freedom of the error sum of squares by the number of missing values.