The basic concepts of experimental design, and completely randomized design

This section discusses the basic concepts of experimental design, and completely randomized design:

. The entire topic is divided into the following subdivisions.

1. Terminologies
2. Principles of Experimentation
3. Completely randomised design (CRD)
4. Least square estimators of the parameters
5. Statistical Analysis
6. ANOVA Table
7. Advantages and disadvantages of CRD
8. Example
9. Conclusion

1. Terminologies

Experiment:

An experiment is a device to obtain answers to some scientific query. It Is a process or study that results in the collection of data. The results of experiments are not known in advance .In comparative experiment we compare the effects of two or more factors on some population characteristics, e.g. comparison of different varieties of crops, different fertilizers in a agricultural experiment ,different medicines in a medical expt.. etc.

Experimental design: Experimental design **is** the process of planning a study to meet the specified objectives. Planning an experiment properly is very important in order to ensure that the right type of data and a sufficient sample size and power are available to answer the research questions of interest as clearly and efficiently as possible.

Experiment units: The smallest subdivision of the experimental material to which the treatments are applied and on which the variable under study is measured is called an experimental units. Thus in an agricultural field experiment, the plot of land on which a treatment is applied is an experimental unit. In a feeding experiment of animals, an animal is an experimental unit.

Treatment: Various objects of comparison in a comparative experiment are called treatments. For eg. In an agricultural experiment, different fertilizers or different varieties of crop are treatments.

Block- Experimental units are collected together to form a relatively homogeneous group. This group is called a Block. A block is also a replicate.

Experimental error: Even when same treatment is applied to different experimental units, the result will vary. A part of this variation is systematic and can be ascribe to different known sources. The other unexplained part of variation is called experimental error and includes all extraneous variations due to inherent variability in the experimental units, errors of measurements and lack of representativeness of the sample to the population of interest.

2. **Principles of Experimentation**: The experimental design involves the three basic principles viz., randomisation, replication and local control

Randomization. The first principle of an experimental design is randomization, which is a random process of assigning treatments to the experimental units. The random process implies that every possible allotment of treatments has the same probability. The purpose of randomization is to remove bias and other sources of extraneous variation, which are not controllable. Hence the treatments must be assigned at random to the experimental units

Replication. The second principle of an experimental design is replication; which is a repetition of the basic experiment. In other words, it is a complete run of all the treatments to be tested in the experiment. In all experiments, some

variation is introduced because of the fact that the experimental units such as individuals or plots of land in agricultural experiments cannot be physically identical. This type of variation can be removed by using a number of trials. We therefore perform the experiment more than once, i.e., we repeat the basic experiment. An individual repetition is called a replicate. The number, the shape and the size of replicates depend upon the nature of the experimental material.

Local Control. The term local control, referring to the amount of balancing, blocking and grouping of the experimental units in to number of homogeneous sub plots. Balancing means that the treatments should be assigned to the experimental units in such a way that the result is a balanced arrangement of the treatments. Blocking means that like experimental units should be collected together to form a relatively homogeneous group.. The main purpose of the principle of local control is to increase the efficiency of an experimental design by decreasing the experimental error. It has been observed that all extraneous sources of variation are not removed by randomization and replication. In other words, we need to choose a design in such a manner that all extraneous sources of variation are brought under control. For this purpose, **Local Control is implemented..**

3. *Completely randomised designs (CRD)*

This is the simplest type of design, based on principles of randomisation and replication. A total of *N* experimental units are available for use in the experiment. These experimental units are as homogeneous as possible; that is no source of variation can be recognized among them under any grouping or arrangement. Suppose that we have p treatments whose effect on the response has to be investigated. The experimental plan is subdivide the N plots randomly in to p parts. Such that j^{th} part consists of n_j plots. ($\sum n_j = N$). The first treatment is allocated to the first set of n_1 plots, the second treatment to the second set of n_2 plots,, etc. The arrangement of N plots in to groups n_1, n_2,…, n_p plots is done in

a completely random manner. Hence the design is called completely randomised design.

The CRD utilises the principles of randomisations and replication in the following way:

Randomisation: n_i experimental units are selected at random and i^{th} treatment is allocated to these experimental units

Replication: Since each treatment appears n_i times.

The data can be tabulated as follows

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	…	y_{1n1}	$y_{1.}$	$\overline{y_{1.}}$
2	y_{21}	y_{22}	…	y_{2n2}	$y_{2.}$	$\overline{y_{2.}}$
.	.	.	…	.	.	.
..	.	.	…	.	.	.
.	.	.	…	.	.	.
p	y_{p1}	y_{p2}	…	y_{pnp}	$y_{p.}$	$\overline{y_{p.}}$

Statistical Analysis is similar to that of one-way classified data,

Represent Y_{ij} as the j^{th} observation taken from treatment i. There will be in general n_i observations under the i^{th} treatment.

Model for the data.: We define the model

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, i=1,2…p, j=1,2,…$n_i$ (4)

Where μ general effect

α_i – i^{th} treatment effect; ϵ_{ij}- error term

4. Least square estimators of the parameters

The parameters μ and α_i are estimated by the method of least squares. i.e. by minimising error sum of squares. L = $\sum_{i=1}^{p}\sum_{j=1}^{n_i}\varepsilon_{ij}^{2} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)^2$

i.e. choose the values of μ and α_i say $\hat{\mu} \quad and \quad \hat{\alpha}_i$ that minimises L.

Which is obtained by solving p+1 simultaneous equations

$\frac{\partial L}{\partial \mu} = 0$ and $\frac{\partial L}{\partial \alpha_i} = 0$, i= 1,2…p_

Differentiating with respect to μ and α_i and equating to zero, we obtain p+1 equations called as normal equations as follows,

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

$$N\mu + \sum_{i=1}^{p} n_i \alpha_i = \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}$$

And

$\frac{\partial L}{\partial \alpha_i} = 0$

$n_i \mu + n_i \alpha_i = \sum_{j=1}^{n_i} y_{ij}$ i= 1,2 …p

These normal equations are not linearly independent, as first equation is equal to the sum of rest p equations. Hence no unique solution exists for μ and $\alpha_{i,\ i=}$ 1,2..p. Since we have defined the treatment effects as deviations from overall mean, we add a independent constraint, $\sum_{i=1}^{p} n_i \alpha_i = 0$ and solve the simultaneous normal equations. Solving we get the solutions as

$\hat{\mu} = \bar{y}_{..}$ and $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ i=1,2..p

The fitted model after substituting the estimates $\hat{\mu}$ and $\hat{\alpha}_i$ in the linear model , we get

$Y_{ij} = \hat{\mu} + \hat{\alpha}_i + \epsilon_{ij}$

Or

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

Or

$Y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$, the error term is chosen that both sides are balanced

Squaring both sides and summing over all the observations we get

$\sum_{i=1}^{p}\sum_{j=1}^{ni}(y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p} n_i(\overline{y_{i.}} - \overline{y_{..}})^2 + \sum_{i=1}^{p}\sum_{j=1}^{ni}(y_{ij} - \overline{y_{i.}})^2$, the cross product vanishes

Or

SST = SSTR + SSE

Where

SST = $\sum_{i=1}^{p}\sum_{j=1}^{ni}(y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p}\sum_{j=1}^{ni} y_{ij}^2 - \frac{y_{..}^2}{N}$ (simplified formula)

SSTR = $\sum_{i=1}^{p} n_i(\overline{y_{i.}} - \overline{y_{..}})^2 = \sum_{i=1}^{p}\frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$

And SSE = $\sum_{i=1}^{p}\sum_{j=1}^{ni}(y_{ij} - \overline{y_{i..}})^2$ computed as SSE = SST-SSTR

Thus total corrected sum of squares can be partitioned into a sum of squares of the differences between the treatment averages and the grand average (SSTR),plus a sum of squares of the differences of the observations within treatments averages(SSE)

5. Statistical Analysis:

We now investigate a formal test of the hypothesis

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$

Against the alternative

H_1: at least one $\alpha_i \neq \alpha_j$ for all i,j

We have assumed that the errors ϵ_{ij} are normally and independently distributes with mean zero and variance σ^2. The observations y_{ij} are normally and independently distributed with mean $\mu + \alpha_i$ and variance σ^2. Thus SST is a sum of squares in normally distributed random variables hence can be shown that SST/σ^2 is distributes as chi-square with N-1 degrees of freedom. Further we can show that SSE/σ^2 is chi-square N-p degrees of freedom and that $SSTR/\sigma^2$ is chi - square variate with p-1 degrees of freedom if the null hypothesis i.e. H_0 : $\alpha_i = 0$ is true. It also implies that $SSTR/\sigma^2$ and SSE/σ^2 independently distributed chi-square random variables. Therefore if the null hypothesis is true, the ratio

$$F_\alpha = \frac{SSTR/(p-1)}{SSE/(N-p)} = \frac{MSSTR}{MSSE} \qquad (1)$$

is distributed as F distribution with p-1 and N-p degrees of freedom.

Equation (1) is the test statistics for testing H_0: there is no differences in treatment means.

6. Analysis of variance Table(ANOVA Table)

Sources of variation	Degree of freedom	Sum of squares	Mean sum of squares	F-Value
Treatment	p-1	SSTR	MSSTR	MSSTR

s			= SSTR/(p -1)	/ MSSE
Error	N-p	SSE	MSSE = SSE/(p-1) (k-1)	
total	N-1	TSS		

We reject H_0 and conclude that there are differences in the treatment means if $F_0 > F_{\alpha, p-1, N-p}$ Where F_0 is computed from equation 1 and $F_{\alpha, p-1, N-p}$, is the table value referring to F table at α level of significance corresponding to p-1and N-p degrees freedom.

7. Advantages and disadvantages of CRD

Advantages:

a) The design is very simple to implement.

b) Any number of treatments can be used with unequal replications and does not make the statistical hypothesis complicated.

c) The statistical analysis is simple. Even if some/all of the observations for any treatment is missing , the statistical analysis of the data does not become complicated

d) The design provides maximum number of degrees of freedom for the estimation of error variance. For small experiments, the precision increases with increasing number of error d.f.

Disadvantage:

1) In most circumstances, the experimental units are not homogeneous, particularly, when a large no. of units involved. The CRD fails to take account of the variation among the experimental units as it does not use the principle of local control, This will increase the value of error variance under CRD

2) CRD is appropriate only if all plots are homogeneous. In reality this will never happen in field experiment, thus the CRD not recommended for field trials.

8. **Example:** A set of data involving four tropical feed stuffs A,B,C and D tested on 22 chicks is given below. All the twenty two chicks are treated alike in all respect except the feeding treatment, Analyse the data.

Data:

Wt. gain in chicks fed on different feeding materials composed of tropical feed stuff						
A:	55	49	42	20	62	73
B:	61	112	30	89	63	
C:	42	97	81	95	92	102
D:	169	137	169	85	154	

Solution:

We have 4 treatments A,B,C and D and we have to compare these 4 treatments on the response vatiable (Wt. Gain). Only the treatments (feed stuffs) affecting the wt. gain other than general effect. Hence we assume the one way anova model

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, i=1,23.4, j=1,2,…$n_i$ (Where n_1-6, n_2-5, n_3=6, n_4=5)

Where μ general effect

α_i – i^{th} treatment effect; ϵij- error term

We test the hypothesis

H_0: $\alpha_1 = \alpha_2 = \alpha_{2=}\alpha_4 = 0$

Against the alternative

H_1: at least one $\alpha_i \neq \alpha_j$, for all i,j, (i≠j) , (i,j: 1,2,3,4)

Step 1:Calculation of correction factor(CF): $\frac{y_{..}^2}{N}$ = $1879^2/22$= 160.0483

Step;2: Calculation of Total sum of squares: $\sum_{i=1}^{p}\sum_{j=1}^{k} y_{ij}^2 - \frac{y_{..}^2}{N}$ =

$(55^2+49^2+\ldots+154^2)$-CF

=37793.318

Step;3: Calculation of Treatment sum of squares= $\sum_{i=1}^{p} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$ =

$(301^2/6+355^2/5+509^2/6+714^2/5)$ –CF = 25583.152

(Total(A) =301, **Total(B) =**355, **Total(C) =509, Total(D) =714)**

Step 4: Calculation of Error sum of squares

ESS = 37793.318-25583.152=12210.167

Anova Table:

Sources of variation	Degree of freedom	Sum of squares	Mean sum of squares	F-Value
Treatments (foodstuffs)	p-1=3	SSTR= 25583.152	MSSTR = SSTR/(p-1)= 8527.717	MSSTR/ MSSE = 12.571
Error	N-p=18	SSE=12210.167	MSSE = SSE/(p-1) (k-1) 678.343	
total	N-1 =21	TSS =37793.318		

Table value for F_{α} is $F_{.05,\ 3,18}$ = 3.16

Conclusion: Sinc $F_{cal}(12.571) > F_{.05,\ 5,15}$ we reject H_{α}. Hence we conclude that all treatment means are not equal

When Ho is rejected we can go for multiple comparison test (method of least square difference or critical difference) if needed to test which pair of treatment are significantly different . Proceed as follows.

Compute $\overline{y_{i.}} - \overline{y_{j.}}$ and $t_{\alpha,n-p}\sqrt{MESS\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$. The Table value is $t_{.05,18}$ is 2.101

Absolute differences OF $\overline{y_{i.}} - \overline{y_{j.}}$ given in the following table

Treatments	A	B	C	D
A	-	25.73333	39.23333*	97.20000*
B	-	-	13.50000	71.46667*
C	-	-	-	57.96667*

Calculation of critical difference $t_{\alpha,n-p}\sqrt{MESS\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

Treatments	A	B	C	D
A	-	34.22	18.51	34.22
B	-	-	34.22	35.34
C	-	-	-	34.22

The mean difference is significant at 5% level for the pairs (A,C), A,D), (B,D), and (C.D) which are the pairs marked *

Conclusion: In this session we have defined some terminologies which are used in design and analysis of experimentation. We have discussed the principles of experimental design. We have introduced completely randomised design and discussed its analysis . We have derived the least square estimates of the parameters of the design. We summarised the complete analysis of CRD using ANOVA table. We also discussed the advantages and disadvantages of CRD. We have discussed the complete analysis of CRD using an example.