

1. INTRODUCTION

1.1 Definition and Brief History of Statistics

- In biology, including agricultural sciences, “the laws of nature” are not that simple.
- Biological phenomena often show variation that obscures the law we want to establish.
- For instance, if we treat two fields in identical way, the yields obtained from the same variety of a crop will not be the same. Similarly, two cows of the same age, breed and body weight fed with the same type and amount of fodder will produce different amount of milk. Thus, variation is a typical feature of biological data.

Statistics: study of numerical data based on variation in nature.

- It is a science, which deals with collection, classification, tabulation, summary and analysis of quantitative data or numerical facts.
- It is the science of creating, developing, and applying techniques such that the uncertainty of inductive inferences may be evaluated.
- The word statistics also refers to numerical and quantitative data such as statistics of births, deaths, marriage, production, yield, etc.

Biometry: The application of statistical methods to the solution of biological problems.

- Biometry is also called biological statistics or bio-statistics.
- The word biometry comes from two Greek roots: 'bios' mean life and metron mean 'to measure'. Thus, biometry literally means the measurement of life.

➤ Events related to statistics

- The term statistics is an old one.
- A state arithmetic technique to assist a ruler who needed to know the wealth and number of his subjects in order to levy a tax or wage a war.
- Gambling, in the form of games of chance, led to the theory of probability originated by Pascal and Fermat
- Introduction of the equation of normal curve was first published in 1733 by de Moivre. De Moivre had no idea of applying his result to experimental observations and his paper remained unknown until Karl Pearson found it in a library in 1924. However, the same result was later developed by two mathematical astronomers, Laplace (1749-1827) and Gauss (1777-1855) independently of one another.
- Charles Darwin (1809-1882), a biologist, received the second volume of Lyell’s book while on the Beagle. Darwin formed his theories later and he may have been stimulated by his reading of this book. Darwin’s work was largely biometrical or statistical in nature and he certainly renewed enthusiasm in biology.
- Gregor Mendel (1822-1884) too, with his studies of plant hybrids published in 1866, had a biometrical or statistical problem.

- In the nineteenth century, the need for a sounder basis for statistics became apparent. Karl Pearson (1857-1936) initially a mathematical physicist applied his mathematics to evolution as a result of the enthusiasm in biology created by Darwin. Pearson spent nearly half a century in serious statistical research. In addition, he founded the journal *Biometrika* and a school of statistics as result the study of statistics gained impetus.

1.2 Definition of Some Basic Terms

Data: Qualitative or quantitative information taken on a certain character. For example, data of height, weight, color, etc.

Variable: A property with respect to which individuals in a sample differ in some way, *e.g.* length, weight, height, color, etc. Characteristics, which show variations, are called random variables.

Variables can be:

1. **Measurement variables:** Variables that can be expressed in a numerical order. They are of two types:
 - a) **Continuous variables:** Variables which can assume an infinite number of values between any two fixed points, *e.g.* height, grain yield, score of students, etc.
 - b) **Discontinuous/Meristic/Discrete variables:** Variables that take any certain fixed numerical values with no intermediate values in between, *e.g.* number of seeds per pod, number of plants in a quadrat, number of students taking the course biometry, etc.
2. **Ranked variables:** Some variables cannot be measured but at least can be ordered or ranked by their magnitude, *e.g.* disease score; no infection(0), mild infection(1), high infection(2), severe infection(3)
 Here we cannot say that the difference between 1 and 2 is identical or equal to the difference between 2 and 3. Such assumption is made for the measurement variables.
3. **Attribute or nominal variables:** Variables that cannot be measured but are expressed qualitatively, *e.g.* color of common bean seed: white, black, red; sex of animals could be male or female, blood group of humans could be A, AB, O, etc.
 When such attribute data are combined with frequency (number of occurrences), they can be treated statistically. Such data are called enumeration data. Suppose we have 18 mixed bean seeds, they can be grouped as: black (3), white (5) and red (10).

Variate (datum): A single reading, score or observation of a given variable. If we measure height of 5 plants from a plot, each of the 5 readings of height will be a variant: 10 cm, 5 cm, 20 cm, 25 cm, 15 cm.

Population and sample

Population: the collection of all the possible observations of interest.

Sample: is a subset from the population selected by a specific procedure.

- In an experiment, we can rarely include the whole population because it is costly, time consuming and sometimes impossible to determine.
- Collection of observations we take from the population.

Sample size: The number of observations in the sample (usually indicated by the symbol n).

Simple random sampling: The basic method of collecting the observations in a sample.

Parameter: A population value, which we generally do not know, but would like to infer (estimate) about. For example, national average yields of maize in 2010 in Ethiopia. Parameters are designated using Greek letters such as μ , Σ , π , etc.

Statistics: Are sample estimates of population value (parameters) and designated using Latin letters such as, s , p , etc.

1.3 Types of Statistics

1. **Descriptive (deductive) statistics:** Are methods, which are used to describe a set of data without involving generalization. Deal with the presentation of research data or any numerical information. Help in summarizing and organizing data so as to make them readable for users, *e.g.* mean, median, mode, standard deviation, etc.
2. **Inferential (inductive) statistics:** It is a statistics, which helps, in drawing conclusion about the whole (population) based on data from some of its parts (samples).

2. STATISTICAL INFERENCE

2.1 Estimation

- Decisions about population is made on the basis of sample information because
 - It is generally difficult and sometimes impossible to consider all the individuals in a population for economic, time and other reasons.
 - For example, we take random sample and calculate the sample mean (\bar{x}) as an estimate of population mean (μ) and sample variance (s^2) as an estimate of population variance (σ^2).

Estimation: A technique of estimating the value of population parameter based on sample values.

Estimator: The formula that is used to make estimation.

Estimate: The resulting sample value.

- There are two kinds of estimations

Point estimation: is a technique by which a single value is obtained as an estimate of a population parameter like $\bar{x} = 10$; $s^2 = 0.6$ are point estimates of μ and σ^2 , respectively.

Interval estimation: is an estimation technique in which two limits within which a parameter is expected to be found are determined. It provides a range of values that might include the parameter with a known probability, *e.g.* confidence intervals. For example, population mean (μ) lies between two points such that $a < \mu < b$ where a and b are lower and higher limits, respectively, and are obtained from sample observations, *e.g.* the average weight of students in a class lies between 50 kg and 60 kg.

2.1.1 Properties of best estimates

- **Unbiasedness**: the expected value of the sample statistic (the mean of its probability distribution) should be equal to the parameter. Repeated samples should produce estimates which do not consistently under or over estimate the population parameter.
- **Consistency**: as the sample size increases then the estimator will get closer to the population parameter. Once the sample includes the whole population, the sample statistic will obviously equal the population parameter.
- **Efficiency**: it has the lowest variance among all competing estimators. For example, the sample mean is a more efficient estimator of the population mean of a variable with a normal probability distribution than the sample median, despite the two statistics being numerically equivalent.
- **Sufficiency**: the estimates should have all the required information about the parameter, which is being estimated. For example, the sample standard deviation mean is a more sufficient estimator than range as it includes all the observations.

2.1.2 Confidence Interval

The sample mean (\bar{x}) and variance (s^2) are point estimates of population mean (μ) and population variance (σ^2), respectively. These sample estimates may or may not be equal to the population mean. Thus, it is better to give our estimation in interval and state that the population mean (μ) is included in the interval with some measure of confidence.

Confidence interval: is a probability statement concerning the limit within which a given parameter lies. Eg. We can say the probability that μ lies within the limit a and b is 0.95 where $a < b$; $P(a < \mu < b) = 0.95$

2.1.3 Estimation of single population mean (μ)

Case 1: When population standard deviation (σ) is known (given), the sample size could be large ($n \geq 30$) or small ($n < 30$)

A $(100 - \alpha) \%$ Confidence Interval for population mean (μ) = $\bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$ where

- α = allowable error rate

- $\bar{x} - Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = L_1$ (lower confidence limit)

- $\bar{x} + Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = L_2$ (upper confidence limit)

- Z is the value of standard normal curve

Example: A certain population has standard deviation of 10 and samples of 49 observations were taken from a population and mean of sample was 1.72. Find point estimate and construct 95% confidence interval for the population mean (μ).

Given: $\sigma=10$; $\bar{x} = 1.72$; $n = 49$. Thus,

- Point estimate = 1.72
- A 95% confidence interval for population mean

$$= \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 1.72 \pm Z_{0.025} \times \frac{10}{\sqrt{49}} = 1.72 \pm 1.96 \times 1.43$$

- $L_1 = 1.72 - (1.96 \times 1.43) = -1.08$
- $L_2 = 1.72 + (1.96 \times 1.43) = 4.52$

Thus, 95% C. I. for population mean is $-1.08 < \mu < 4.52$. This means, the probability that the population mean lies in between -1.08 and 4.52 is 0.95 or we are 95% confident that the population mean is included in this range.

We increase the confidence coefficient from 0.95 to 0.99 , the width of the interval increases and the more certain we can be that the true mean is included in this estimated range. However, as we increase the confidence level, the estimate will lose some precision (will be more vague).

Case 2: When σ (pop standard deviation) is unknown

In many practical situations it is difficult to determine the population mean and the population standard deviations. Under such cases, a sample standard deviation (s) can be used to approximate the population standard deviation (σ).

A). Estimation with large sample size ($n \geq 30$)

A $(100 - \alpha) \%$ Confidence Interval for population mean (μ) = $\bar{x} \pm Z_{\alpha/2} \times \frac{s}{\sqrt{n}}$

Example: A poultry scientist is interested in estimating the average weight gain over a month for 100 chicks introduced to a new diet. To estimate the average weight gain, he has taken a random sample of 36 chicks and measured the weight gain, the mean and standard deviation of this sample were $\bar{x} = 60$ g and $s = 5$ g.

Find the point estimate and construct 99% confidence interval for population mean.

- Point estimate = 60 g
- A 99% C. I. = $\bar{x} \pm Z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 60 \pm Z_{0.005} \times \frac{5}{\sqrt{36}} = 60 \pm 2.58 \times 0.83$

Thus, 99% C.I. for population mean = $57.86 \text{ g} < \mu < 62.14 \text{ g}$. The probability that mean weight gain per month is between 57.86 g and 62.14 g is 0.99 .

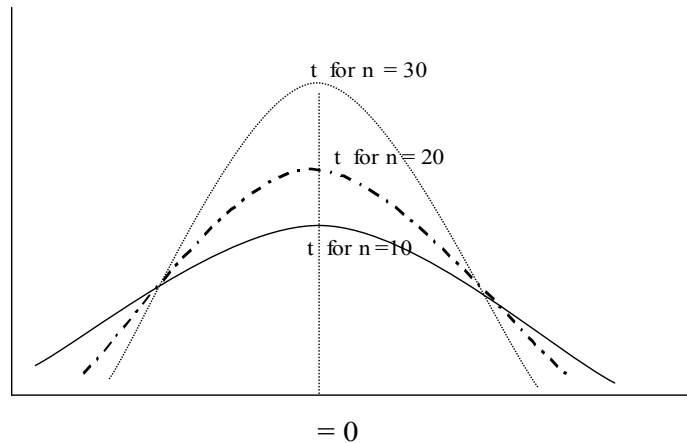
B) Estimation with small sample size ($n < 30$)

The central limit theorem states that as sample size increases, the sampling distribution approaches a normal distribution. However, samples of small size do not follow the normal distribution curve, but t-distribution.

Thus, A $(100 - \alpha)\%$ C.I. for population mean (μ) = $\bar{x} \pm t_{\alpha/2} (n-1) \times \frac{s}{\sqrt{n}}$ where $n-1$ is degree of freedom:

Properties of t-distribution:

- symmetrical in shape and has a mean equal to zero like normal (Z) distribution.
- the shape of t-distribution is flatter than the Z distribution, but as sample size approaches 30, the flatness associated with the t-distribution disappears and the curve approximates the shape of Z-distribution.
- there are different distributions for each degree of freedom



Example: A sample of 25 horses has average age of 24 years and standard deviation of 4 years. Calculate 95% C.I. for population mean.

A 95% C.I. = $\bar{x} \pm t_{\alpha/2} (n-1) \times \frac{s}{\sqrt{n}} = 24 \pm t_{0.025} (25-1) \times \frac{4}{\sqrt{25}} = 24 \pm 2.064 \times 0.8$. Thus, a 95% C.I. = 22.35 years < μ < 25.65 years

2.1.4 Estimating the difference between two population means

There are three cases:

Case 1: When σ_1^2 & σ_2^2 (population variances) are known (n_1 and n_2) could be large or small)

- Point estimate = $\bar{x} - \bar{y}$ if $\bar{x} > \bar{y}$ or $\bar{y} - \bar{x}$ if $\bar{y} > \bar{x}$
- A $(100-\alpha)\%$ C.I. = $\bar{x} - \bar{y} \pm Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ if $\bar{x} > \bar{y}$ or $\bar{y} - \bar{x} \pm Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ if $\bar{y} > \bar{x}$

where \bar{x} and \bar{y} are sample mean of 1st and 2nd group; n_1 and n_2 are sample sizes of 1st and 2nd group; σ_1^2 & σ_2^2 are population variances of 1st and 2nd group.

Case 2: σ_1^2 & σ_2^2 are unknown, but with large sample sizes (n_1 and $n_2 \geq 30$)

For large sample size, s_1^2 is an estimate of σ_1^2 and s_2^2 is an estimate of σ_2^2

- Point estimate = $\bar{x} - \bar{y}$ if $\bar{x} > \bar{y}$ or $\bar{y} - \bar{x}$ if $\bar{y} > \bar{x}$

$$- A(100-\alpha) \% \text{ C.I.} = \bar{x} - \bar{y} \pm Z_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ if } \bar{x} > \bar{y} \text{ or } \bar{y} - \bar{x} \pm Z_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ if } \bar{y} > \bar{x}$$

Example: The performance of two breeds of dairy cattle (Holstein & Newjersy) was tested by taking a sample of 100 cattle each. The daily milk yield was as follows. Find point estimate and construct a 90% C. I. for population mean difference.

- Holstein (x): $\bar{x} = 15$ liters; $s_1 = 2$ liters.
- Newjersy(y): $\bar{y} = 10$ liters; $s_2 = 1.5$ liters
- Point estimate = $\bar{x} - \bar{y} = 15 - 10 = 5$ liters
- A 90% C.I. = $\bar{x} - \bar{y} \pm Z_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 15 - 10 \pm Z_{0.05} \times \sqrt{\frac{(2)^2}{100} + \frac{(1.5)^2}{100}} = 5 \pm 1.65 \times 0.25$

Thus, a 90% C.I. for population mean difference is $4.59 \text{ lt} < \mu_1 - \mu_2 < 5.41 \text{ lt}$

Case 3: σ_1^2 & σ_2^2 are unknown and n_1 and n_2 are small (< 30)

- Point estimate = $\bar{x} - \bar{y}$ if $\bar{x} > \bar{y}$ or $\bar{y} - \bar{x}$ if $\bar{y} > \bar{x}$
- A $(100-\alpha)\%$ C.I. for population mean difference

$$= \bar{x} - \bar{y} \pm t_{\alpha/2} (n_1+n_2-2) \times sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$Sp (\text{pooled standard deviation}) = \sqrt{\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1 + n_2 - 2}}$$

n_1, n_2 are sample sizes; $n_1 + n_2 - 2$ = degrees of freedom

Example: Two groups of nine framers were selected, one group using the new type of plough and another group using old type of plough. Assume the soil, weather, etc. conditions are the same for both groups (variation is only due to plough types). The yields and standard deviation were:

$$\bar{x} = 35.22 \text{ qt}; s_1 = 4.94 \text{ qt}; \bar{y} = 31.55 \text{ qt}; s_2 = 4.47 \text{ qt}$$

Estimate the difference between two plough types and construct a 99% confidence interval for the difference of the population means:

- Point estimate = $\bar{x} - \bar{y} = 35.22 - 31.55 = 3.67 \text{ qt}$
- A 99% C. I. = $35.22 - 31.55 \pm t_{0.005} (9+9-2) \times sp \sqrt{\frac{1}{9} + \frac{1}{9}}$

$$Sp (\text{pooled standard deviation}) = \sqrt{\frac{(9-1) \times (4.94)^2 + (9-1) \times (4.47)^2}{9+9-2}} = 3.67 \pm 6.48$$

Thus, 99% C.I. for pop mean difference = $-2.81 \text{ qt} < \mu_1 - \mu_2 < 10.15 \text{ qt}$

2.2 Hypothesis Testing

2.2.1 Types of hypothesis and errors in hypothesis testing

In attempting to reach at decisions, we make assumptions or guesses about the population. Such assumption, which may or may not be true is called statistical hypothesis.

Examples:

- The ratio of male to female in Ethiopia is 1:1
- The cross of 2 heterozygous varieties for color of tomato will produce plants with red and white flowers in the ratio of 3:1. $Rr \times Rr$ $RR: 2Rr: rr$ where red is dominant over white.

There are two types of hypothesis:

1. Null hypothesis (H_0): Is a hypothesis which states that there is no real difference between the true values on the population from which we sampled, *e.g.* male to female ratio is 1:1.
2. Alternate hypothesis (H_A): Any hypothesis which differs from a given null hypothesis
 - H_A : sex ratio is not 1:1
 - Male ratio > Female ratio
 - Male ratio < Female ratio

If we find that results observed in random sample differ markedly from those expected under null hypothesis on the basis of pure chance, we will say the observed differences are significant and we would reject the null hypothesis.

A procedure that enables us to decide whether to accept or reject the null hypothesis or procedure used to determine whether observed samples differ significantly from expected results are called test of hypothesis or rules of decision.

Rejection and Acceptance of Hypothesis

A hypothesis is rejected if the probability that it is true is less than some predetermined probability. The predetermined probability is selected by the investigator before he collects his data based on: his research experience, the consequences of an incorrect decision and the kind of risk the researcher prepares to take.

When the null hypothesis is rejected, the finding is said to be statistically significant at that specified level of significance. When the available evidence does not support the rejection of the null hypothesis, the finding is said to be non-significant.

Errors in Hypothesis Testing

Two kinds of errors are encountered in hypothesis testing:

- Type I error (α) is committed when a true null hypothesis is rejected. Concluding there is a difference between the groups being studied when, in fact, there is no difference. Thus, it would seem reasonable to select a small value of α .
- Type II error (β) committed when a false null hypothesis is accepted. Concluding there is no difference between the groups being studied when, in fact, there is a difference.

The two kinds of correct decisions are accepting a true null hypothesis and rejecting a false null hypothesis.

The probability of type I error is the significance level, *i.e.* in a given hypothesis, the maximum probability which we will be willing to risk a type I error is called the level of significance of the test. It is denoted by α and is generally specified before samples are taken. In practice 5% or 1% levels of significance are in common use. A 5% level of significance means that there are about 5 chances in 100 that we would reject null hypothesis (H_0) when it should be accepted, *i.e.* we are 95% sure that we make a correct decision.

The probability of type II error (accepting a false null hypothesis) is represented by β . Type II error usually occurs when the sample size is small. The more an investigator protects his experiment against type I error, the greater the likelihood of committing type II error.

Thus, the value of α must be chosen to minimize type I as well as type II error.

Power of Test: This is defined as the probability of rejecting the hypothesis when it is false and symbolically given as $(1-\beta)$. We have to design our experiment in such away that the power of test is maximized.

2.2.2 One- tailed and two-tailed tests

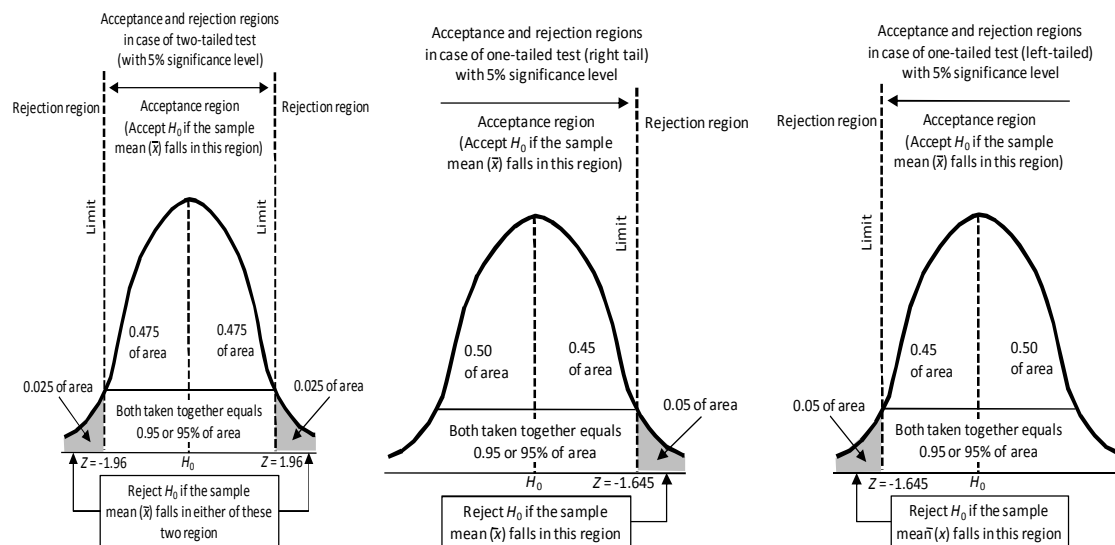
In most cases, the H_0 (null hypothesis) is one of no effect (e.g. no difference between two means) and the H_A (the alternative hypothesis) can be in either direction; the H_0 is rejected if one mean is bigger than the other mean or vice versa. This is termed a two-tailed test because large values of the test statistic at either end of the sampling distribution will result in rejection of H_0 .

a. In two-tailed test, to do a test with $\alpha = 0.05$, then we use critical values of the test statistic at **$\alpha/2 = 0.025$** at each end of the sampling distribution.

b. Sometimes, our H_0 is more specific than just no difference. We might only be interested in whether one mean is bigger or smaller than the other mean but not the other way. In one-tailed test, to do a test with **$\alpha = 0.05$** , then we use critical values of the test statistic at $\alpha = 0.05$ at one end of the sampling distribution.

Thus, in one tailed test, the area corresponding to the level of significance is located at one end of the sampling distribution while in a two tailed test the region of rejection is divided into two equal parts, one at each end of the distribution.

Most statistical tables either provide critical values for both one- and two-tailed tests but some just have either one- or two-tailed critical values depending on the statistic, so make sure that you look-up the correct α values when you use tables. Statistical software usually produces two-tailed α values so you should compare the α value to 0.10 for a one-tailed test at 0.05.



5.2.3 Test of a single population mean (μ)

Procedure:

- State the null and alternate hypothesis
- Determine the significance level (α). The level of significance establishes a criterion for rejection or acceptance of the null hypothesis (10%, 5%, 1%)
- Compute the test statistic: The test statistic is the value used to determine whether the null hypothesis should be rejected or accepted. The test statistic depends on sample size or whether the population standard deviation is known or not:
 - when σ is known n could be large or small, the test statistic

$$Z_c = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}} \text{ (normal distribution)}$$

- when σ^2 unknown, but with large sample size ($n \geq 30$)

$$Z_c \text{ (test statistic)} = \frac{(\bar{x} - \mu)}{s / \sqrt{n}} \text{ (normal distribution)}$$

- when σ unknown and with small sample size ($n < 30$)

$$t_c \text{ (test statistic)} = \frac{(\bar{x} - \mu)}{s / \sqrt{n}} \text{ (t-distribution)}$$

- Determine the critical regions (value). The critical value is the point of demarcation between the acceptance or rejection regions. To get the critical value, read Z or t -table depending on the test statistic used.
- Compare calculated value (test statistic) with table value of Z or t and make decision. If the absolute value of calculated value (test statistic) is greater than table Z or t -value, reject the null hypothesis and accept the alternate hypothesis.

- If $|Z_c|$ or $|t_c| > Z$ or t table value at the specified level of significance, reject the H_0 (null hypothesis)
- If $|Z_c|$ or $|t_c| \leq Z$ or t -table value at the specified level of significance, accept H_0

Example 1: A standard examination has been given for several years with mean (μ) score of 80 and variance (σ^2) of 49. A group of 25 students were taught with special emphasis on reading skills. If the 25 students obtained a mean grade of 83 on the examination, is there reason to believe that the special emphasis changed the result on the test at 5% level of significance assuming that the grades are normally distributed?

$\mu = 80$; $\sigma^2 = 49$; $\bar{x} = 83$; $n = 25$; $\alpha = 0.05$

1. State the null and alternate hypothesis

$H_0: \mu = 80$

$H_A: \mu \neq 80$ (two-tailed test)

2. Compute the test statistic: we use the normal (Z-distribution) because of known population variance

$$Z_c = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}} = \frac{(83 - 80)}{7 / \sqrt{25}} = 2.14$$

3. Obtain the table Z-value at $Z_{\alpha/2} = Z_{0.025} = 1.96$

4. Make decision. Since $|Z_c|$ (2.14) $>$ Z-table (1.96), we reject the null hypothesis and accept the alternate hypothesis, *i.e.* there is sufficient reason to doubt the statement that the mean is 80.

Example 2: It is known that under good management Zebu cows give an average milk yield of 6 liters/day. A cross was made between Zebu and Holstein and the result from a sample of 49 cross breeds gave average daily milk yield of 6.6 liters with standard deviation of 2 liters. Do the cross breeds perform better than the pure Zebu cows at 5% level of significance?

Given: $\mu = 6$ liters; $n = 49$; $\bar{x} = 6.6$; $s = 2$

1. State the null and alternate hypothesis

$H_0: \mu = 6$ (No difference between cross breed and pure Zebu)

$H_A: \mu > 6$ (Cross breeds perform better than the pure zebu). This is one-tailed test.

2. Compute the test statistic: we use the normal (Z-distribution) because of large sample size.

$$Z_c = \frac{(\bar{x} - \mu)}{s / \sqrt{n}} = \frac{(6.6 - 6)}{2 / \sqrt{49}} = 2.1$$

3. Obtain the table Z-value: $Z_{0.05} = 1.65$

4. Make decision. Since $|Z_c|$ (2.1) $>$ Z-table (1.65), we reject the null hypothesis and accept the alternate hypothesis, *i.e.* cross breeds perform better than the pure Zebu or cross breeds gave statistically significant higher milk yield than the pure Zebus.

2.2.4 Test of the difference between two means

Procedure:

1. State the null and alternate hypothesis

2. Compute the test statistic

– **Case I:** When σ_1^2 & σ_2^2 (population variances) are known (n_1 and n_2) could be large or small

$$Z_c = (\bar{x} - \bar{y}) / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

– **Case II:** When σ_1^2 & σ_2^2 are unknown, n_1 & n_2 are large (≥ 30)

$$Z_c = (\bar{x} - \bar{y}) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

– **Case III:** When σ_1^2 & σ_2^2 unknown, n_1 & n_2 small (< 30);

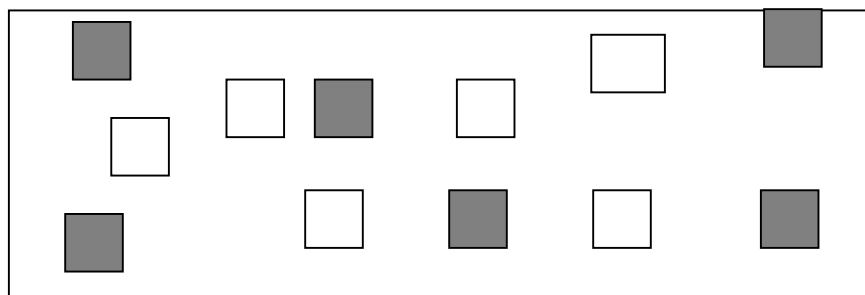
$$t_c = (\bar{x} - \bar{y}) / sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{where} \quad sp \text{ (pooled standard deviation)} = \sqrt{\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}}$$

3. Read Z or t- table depending on the test-statistic used

4. Make decision. If $|Z_c|$ or $|t_c| > z$ table or t table, reject the null hypothesis at the specified level of significance.

Independent t-test

In this case, the allocation of treatments on experimental units is done completely at random, e.g. varieties A (filled) & B (open) are allocated to 12 fields, each on 6 fields randomly.



Example: A researcher wanted to compare the potentials of 2 different types of fertilizer. He conducted an experiment at 10 different locations and the two fertilizer types were randomly applied in five of the fields each on maize. The following yields were obtained in tons/ha.

Fertilizer x:	9.8	10.6	12.3	9.7	8.8
Fertilizer y:	10.2	9.4	9.1	11.8	8.3

Is there any difference in potential yield effect of the fertilizers at 5% level of significance?

Solution

1. Calculate the means and standard deviations: $\bar{x} = 10.24$; $s_1 = 1.32$; $\bar{y} = 9.76$; $s_2 = 1.33$; n_1 & $n_2 = 5$

2. State the hypothesis

$$H_0: \mu_1 = \mu_2;$$

$$H_A: \mu_1 \neq \mu_2 \text{ (two tailed test)}$$

3. Compute the test statistic (σ_1^2 & σ_2^2 unknown, n_1 & n_2 small)

$$tc = (\bar{x} - \bar{y})/sp \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (10.24 - 9.76)/sp \times \sqrt{\frac{1}{5} + \frac{1}{5}}$$

$$sp = \sqrt{\frac{(5-1)(1.32)^2 + (5-1)(1.33)^2}{5+5-2}} = 1.32$$

$$= (10.24 - 9.76)/1.32 \sqrt{\frac{1}{5} + \frac{1}{5}} = 0.58$$

4. Read t- table value for $t_{(\alpha/2)}$ at $5 + 5 - 2$ d.f., $t_{0.025}(8) = 2.306$.

5. Make decision. Since t-calculated (0.58) < t-table (2.306), we accept the null hypothesis, *i.e.* the data do not give sufficient evidence to indicate the difference in potential yield of two fertilizers. Thus, the difference between the two fertilizer types is not statistically significant.

Matched (paired) t-test

Paired (matched) data occur in pairs, *e.g.* blood pressure measure of patient before medication compared to blood pressure measure after medication of the same person.

Example: An experiment was done to compare the yields (qt/ha) of two varieties of maize, data were collected from seven farms. At each farm, variety A was planted on one plot and variety B on the neighboring plot.

Farm:	1	2	3	4	5	6	7
Variety A:	82	68	109	95	112	76	81
Variety B:	88	66	121	106	116	79	89

Test whether there is a significant difference between two varieties at 5% level of significance.

Solution

1. Calculate the differences, mean of the differences, and variance of the differences as shown below

Farm:	1	2	3	4	5	6	7	Σ
Variety A:	82	68	109	95	112	76	81	
Variety B:	88	66	121	106	116	79	89	
Difference (d)	-6	2	-12	-11	-4	-3	-8	-42
$(d - \bar{d})^2$	0	64	36	25	4	9	4	142

$$\bar{d} \text{ (mean of the difference)} = \frac{\sum d}{n} = \frac{-42}{7} = -6$$

$$S_d^2 \text{ (variance of the difference)} = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{142}{7-1} = 23.67$$

Note that in paired data some farms are poor (2, 6), thus low yields for both varieties are obtained while others are good (3, 5) and thus high yield for both varieties.

2. State the hypothesis: $H_0: \mu_1 = \mu_2$; $H_A: \mu_1 \neq \mu_2$ (two-tailed test)

3. Calculate the test statistic

$$t_c = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = \frac{-6}{\sqrt{\frac{23.67}{7}}} = \frac{-6}{1.84} = -3.26$$

4. Read table t value: $t_{\alpha/2} (n-1) = t_{0.025}(n-1) \text{ d.f.} = t_{0.025}(6) = 2.447$

5. Make decision: Since $|t_c| (3.26) > t\text{-table} (2.447)$, there is a significant difference between the varieties of maize at 5% level of significance.

3. PRINCIPLES OF EXPERIMENTAL DESIGN

3.1 Introduction

Research: can be broadly defined as systematic investigation in to a subject to discover new facts or principles or to confirm or deny the results of previous finding.

- In research, a scientist identifies solution to problems through experimentation
- Such investigation will help in decision making such as recommending a new procedure, a new fertilizer rate, a new pesticide, etc.

Experiment: It is a planned inquiry to obtain new facts or to confirm or deny the results of previous experiments.

- It is a device or means of getting an answer to the problem under consideration.
- Experiment is an important tool of research.

The procedure for research is generally known as the scientific method, which, although difficult to define precisely, it usually involves, the following steps:

- a. Formulation of hypothesis: a tentative explanation or solution
- b. Planning an experiment to objectively test the hypothesis
- c. Careful observation and collection of the data
- d. Analysis and interpretation of the experimental results.

➤ Some important characteristics of well-planned experiments are:

- a. **Simplicity:** The selection of treatments and the experimental arrangement should be as simple as possible, and consistent with the objectives of the experiment.
- b. **Degree of precision:** The probability should be high so that the experiment will be able to measure differences with the degree of precision the experimenter desires. This implies an appropriate design and sufficient replication.
- c. **Absence of systematic error:** The experiment must be planned to ensure that experimental units receiving one treatment in no systematic way differ from those receiving another treatment so that an unbiased estimate of each treatment effect can be obtained.
- d. **Range of validity of conclusions:** Conclusions should have as wide range of validity as possible. An experiment replicated in time and space would increase the range of validity of the conclusions that could be drawn from it. A factorial set of treatments is another way for increasing the range of validity of an experiment. In a factorial experiment, the effects of one factor are evaluated under varying levels of a second factor.
- e. **Calculation of degree of uncertainty:** In any experiment there, is always some degree of uncertainty as to the validity of the conclusions. The experiment should be designed so that it is possible to calculate the probability of obtaining the observed results by chance alone.

3.2 Design of Experiments

The term refers to five interrelated activities required in the investigation. These are:

- a. Formulating statistical hypothesis and making plans for laying out, collection and analysis of data

- b. Stating the decision rules to be followed in testing statistical hypothesis
- c. Collecting data according to plan
- d. Analyzing data according to plan
- e. Making decisions based on decision rules

Purposes of experimental designs:

- a. To provide estimates of a treatment effects or differences among treatment effects
- b. To provide an efficient way of testing hypothesis about the response to treatments
- c. To assess the reliability of estimates and assumptions
- d. To estimate the variability of the experimental material
- e. To increase precision by eliminating extraneous external source of variation from the comparisons of interest
- f. To provide a systematic, and efficient pattern of conducting an experiment

3.3 Concepts Commonly Used in Experimental Design

Treatment: It is an amount of material or a method that is to be tested in the experiment such as crop varieties, insecticides, feedstuffs, fertilizer rates, method of land preparation, irrigation frequency, etc.

Experimental unit: It is an object on which the treatment is applied to observe an effect, *e.g.* cows, plot of land, petri-dishes, pots, etc

- ❖ Effect of different rations on milk production, ration is a treatment and animal is the experimental unit
- ❖ Study of different fertilizer rates on yield of maize, the fertilizer rates are treatment and plot of land is experimental unit.

Experimental error: It is a measure of the variation, which exists among observations on experimental units treated alike. Variation generally comes from two main sources:

1. Inherent variability that exists in the experimental material to which treatments are applied.
2. Lack of uniformity in the physical conduct of an experiment or failure to standardize the experimental techniques such as lack of accuracy in measurement, recording data on different days, etc.

Methods aimed at reducing the experimental error

- a) Increase the size of experiment either through provision of more replicates or by inclusion of additional treatments
- b) Refine the experimental technique
 - Have uniformity in the application of treatments such as equally spreading of fertilizers, recording data on the same day, etc.
 - Control should be done over external influences so that all treatments produce their effects under comparable conditions, *e.g.* protect against diseases, insects, etc. as their effects are not uniform on all plots.

c) Blocking: Dividing the field into several homogenous parts. Blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of analysis of variance.

Replication: A situation where a treatment appears more than once in an experiment, it is said to be replicated. The functions of replication are:

- a) It provides an estimate of experimental error because it provides several observations on experimental units receiving the same treatment. For an experiment on which each treatment appears only once, no estimate of experimental error is possible and when there is no method of estimating the experimental error, there is no way to determine whether observed differences indicate the real differences or due to inherent variability.
- b) It improves the precision of an experiment: As the number of replicates increase, the estimates of population means as observed treatment means become closer to the true value.
- c) It increases the scope of inference and conclusion of the experiments: Field experiments are normally repeated over years and locations because conditions vary from year to year and location to location. The purpose of replication in space and time is to increase the scope of inference. The results of an experiment are applicable only to conditions that are similar to that condition.

➤ **Factors determining the number of replications are:**

- a) The degree of precision required: the higher the precision desired the greater is the number of replicates (the smaller the departure from null hypothesis to be measured or detected, the greater the number of replicates)
- b) Variability of experimental units: certain experimental materials are more variable than others. For the same precision, less replication is required on uniform experimental units (soils) than on variable experimental unit (soils).
- c) The number of treatments: more number of replications are needed with few treatments than with many treatments
- d) The type of experimental designs also affect the precision of an experiment and the required number of replications, *e.g* in Latin Square Design the number of replications should be equal to the number of treatments while in balanced lattice design the number of replications is one more than the square root of the number of treatments, *i.e.* $r = \sqrt{t} + 1$ where r is number of replications and t is number of treatments.
- e) Fund and time availability also determines the number of replicates: If there are adequate fund and time, more replications can be used.

❖ Usually three replications are taken as the minimum number for standard experiments

➤ Two procedures for calculating the number of replications are described below:

Procedure 1: This method takes into consideration the number of variability of experimental material and field, which is measured in terms of coefficient of variation (CV) and standard error of means (SEM). To calculate the number of replications we can use the following formula:

$$N = (CV/SEM)^2$$

Where N = Number of replications; CV = Coefficient of variation; SEM = Standard Error of Means.

Example: $CV (\%) = 12$; $SEM \pm = 6$; $N = (12/6)^2 = 4$

Procedure 2: If CV & SEM are not known, then the number of replications can be arrived at using the principle that the precision of treatment comparisons increases if the experimental error is kept to minimum. The experimental error can be kept to the minimum by providing more degrees of freedom for the experimental error. In other words, a lower number of degrees of freedom for experimental error results in enlarged experimental error. Based on this principle, the number of degrees of freedom for error should not be less than 15 (not less than 10 in any case).

When 't' treatments are replicated 'r' times, the error is based on (t-1) (r-1) degrees of freedom in Randomized Complete Block Design (RCBD) and t(r-1) in Completely Randomized Design (CRD), which should not be less than **15**.

No. of treatments:	2	3	4	5	6
Minimum no. of replications (CRD):	9	6	5	4	4
Minimum no. of replications (RCBD):	16	9	6	5	4

Increasing either number of replications or plot size can improve precision, but the improvement achieved by doubling plot size is almost always less than the improvement achieved by doubling replications.

Randomization: Assigning the treatments to the experimental units in such away that any unit has equal chance to receive any treatment, *i.e.* every treatment should have an equal chance of being assigned to any experimental units. Thus, a particular treatment should not be consistently favored or disfavored.

Purposes of randomization

- To eliminate bias: randomization ensures that no treatment is favored or discriminated against the systematic assignment to units in a design
 - To ensure independence among the observations. This is necessary to provide valid significance tests.
- Randomization is usually done by using tables of random numbers or by drawing cards or lots.

Confounding: It occurs when the differences due to experimental treatments, *i.e.* the contrast specified in your hypothesis, cannot be separated from other factors that might be causing the observed differences.

Example, if you wished to test the effect of a particular hormone on some behavioral response of sheep. You create two groups of sheep, males and females, and inject the hormone into the male sheep and leave the females as the control group. Even if other aspects of the design are ok, differences between the means of the two groups cannot be definitely attributed to effects of the hormone alone. The two groups are also different in gender and this may also be, at least partly, determining the behavioral responses of the sheep. In this example, the effects of hormone are confounded with the effects of gender.

Control: It is a part of the experiment that is not affected by the factor or factors studied, but otherwise encounter exactly the same circumstances as the experimental units treated with the investigated

factor(s). For example, when investigating the effect of spraying micronutrients, the crop being in the control should also be sprayed with the same amount of water except the micro nutrients. The reason for this way to work is to make sure that it is only the effect of the substance of interest that is investigated. Otherwise, it may not be possible to draw conclusions of the reason(s) to the outcome of the experiment.

Local Control: The principle of local control is another important principle of experimental designs. Under it the extraneous factor, the known source of variability, is made to vary deliberately over as wide a range as necessary and these needs to be done in such a way that the variability it causes can be measured and hence eliminated from the experimental error. In other words, according to the principle of local control, we first divide the field into several homogeneous parts, known as blocks, and then each such block is divided into parts equal to the number of treatments. Then the treatments are randomly assigned to these parts of a block. Dividing the field into several homogenous parts is known as ‘blocking’. In general, blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of analysis of variance. In brief, through the principle of local control we can eliminate the variability due to extraneous factor(s) from the experimental error.

Degrees of Freedom (D.F.): is the number of observations in a sample that are “free to vary” after knowing the mean when determining the variance. After determined the mean, then only $n-1$ observations are free to vary because knowing the mean and $n-1$ observations, the last observation is fixed. A simple example – say we have a sample of observations, with values 3, 4 and 5. We know the sample mean (4) and we wish to estimate the variance. Knowing the mean and one of the observations does not tell us what the other two must be. But if we know the mean and two of the observations (e.g. 3 and 4), the final observation is fixed (it must be 5). So, knowing the mean, only two observations ($n-1$) are free to vary. As a general rule, the d.f. is the number of observations minus the number of parameters included in the formula for the variance.

3.4 Analysis of Variance

Analysis of variance: It is defined as an arithmetic technique whereby the total variation in a set of data is partitioned into several components.

- Analysis of variance was first introduced by R.A Fisher in 1930s.
- Analysis of variance is used in all fields of research where data are quantitatively measured and it is used: to estimate and test about population variance and to estimate and test about population means

3.4.1 General procedures in analysis of variance

- State the model: symbolic representation that describes data under consideration
- State the hypothesis
- Make calculations
- Construct analysis of variance table, which is used to summarize the calculations in table form for quick assessment of the results
- Make decision: decide either to accept or reject the hypothesis

3.4.2. General assumptions underlying the analysis of variance

I. The treatment and replication effects should be additive

The effect of treatment for all replications and the effect of replication for all treatments should remain constant.

A hypothetical set of data with additive and multiplicative effect of treatments and replications.

Additive effect

Treatment	Replication		Replication effect
	I	II	(I-II)
A	180	120	60
B	160	100	60
Treatment effect (A-B)	20	20	

- Note that the effect of treatments is constant over replications and the effect of replications is constant over treatments.

Multiplicative effect

Treatment	Replication		Replication effect	Log ₁₀		Treatment effect
	I	II	(II-I)	I	II	
A	10	20	10	1.00	1.30	0.30
B	30	60	30	1.48	1.78	0.30
Treatment effect (B-A)	20	40		0.48	0.48	

- Here, the treatment effect is not constant over replications and the effect of replications is not constant over the treatments.
- The multiplicative effects are often encountered in experiments designed to evaluate the incidence of diseases and insects. This happens because the changes in insect and disease incidence usually follow a pattern that is in multiple of the initial incidence. When effects are multiplicative, the logarithmic transformations of the data show the additive effect. Thus, in such cases conduct the analysis and

mean separation using the transformed data, and in tables present transformed means in parenthesis alongside their back transformed values out of parenthesis.

II. Experimental errors (residuals) must be independent

The experimental error of one treatment should not be related to or dependent upon that of another treatment. One example, where residuals may be dependent is in the study of pesticides: a high level of pesticide on one plot of land may spread to adjacent plots, thereby changing the conditions on the following plots. The assumption of independence of errors is usually assured by the use of proper randomization, *i. e* treatments are applied (assigned) at random to experimental units. However, in systematic designs where the treatments are assigned systematically instead of randomly, the assumption of independence of errors is usually violated. The simplest way to detect independence of error is to check the experimental layout if it is random or not.

III. Experimental errors (variances) must be homogeneous (homoscedasticity)

The variances of the treatments must be the same. When some treatments have errors that are exceptionally higher or lower it is called heteroscedasticity. First type of heterogeneity of variances is usually associated with data whose distribution is not normal. Count data, such as the number of infected plants per plot usually follow a Poisson distribution where the variance equals to the mean ($s^2 = \bar{x}$).

- Heterogeneity of variances occurs usually when some treatments have errors that are exceptionally higher or lower than others.

IV. Experimental errors are normally distributed

It can be assumed that the observations within each ‘group’ or treatment combinations come from a normally distributed population.

Data such as number of infested plants per plot usually follow Poisson distribution and data such as percent survival of insects or percent plants infected with a disease assume the binomial distribution.

4. Completely Randomized Design (CRD)

4.1. Uses, advantages and disadvantages

Uses:

1. It is useful when the experimental units (plots) are essentially homogeneous and where environmental effects are relatively easy to control, *e.g.* laboratory and greenhouse experiments. For field experiments where there is generally larger variation among experimental plots like in soil fertility, slope, etc. the CRD is rarely used.
2. It is useful if we suspect that large fraction of the units may not respond or may be lost during the experiment because it is easy to handle missing data in Analysis of Variance unlike in other designs.
3. It is useful for experiments in which the total number of experimental units is limited, because it provides maximum degrees of freedom for error

Advantages

1. It is flexible in that the number of treatments and replications can vary, *i.e.* the number of replications need not be the same from one treatment to another
2. The statistical analysis is simple even with unequal replications and it is not complicated by loss of data or missing observations
3. Loss of information due to missing data is small as compared to other designs
4. The design provides the maximum degree of freedom for estimating the experimental error. This improves the precision of the experiment and is important with small experiments where degrees of freedom for experimental error are less than 20.

Disadvantage:

The main objection to the CRD is that it is often inefficient as there is no way of controlling the experimental error. Since randomization is unrestricted the experimental error includes the entire variation over the experimental units except that due to treatment.

4.2 Randomization and layout

In this design, treatments are assigned to the experimental units completely at random.

Assume that we want to do a pot-experiment on the effect of inoculation of 6-strains of rhizobia on nodulation of common bean using five replications.

Randomization can be done by using either lottery method or table of random numbers.

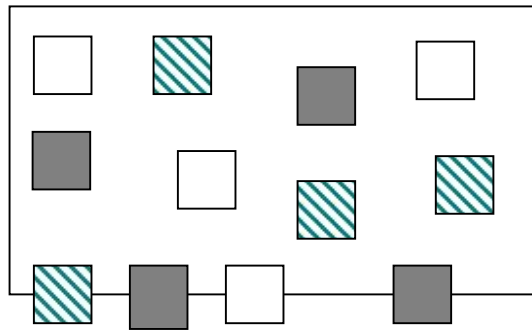
A. Lottery Method

1. Arrange 30 pots of equal size filled with the same type of soil and assign numbers from 1 to 30 in convenient order.
2. Obtain 30 identical slips of paper, label 5 of them with treatment A, 5 of them with treatment B, with C, with D, with E and with F (6 treatments \times 5 replications). Place the slips in box or hat, mix thoroughly and pick a piece of paper at random, the treatment labeled on this paper is assigned to unit 1 (pot 1), without returning the first slip to box, select another slip and the treatment named on this slip is assigned to unit 2 (pot 2) and continue this way until all 30 slips of paper have been drawn.

B. Use of Table of Random Numbers

1. Arrange 30 pots of equal size filled with the same type of soil and assign numbers from 1 to 30 in convenient order.
2. Locate starting point in table of random numbers by closing your eyes and pointing a finger to any position of random number
3. Moving up to down or right to left from the staring, record the first 30 three digit random numbers in sequence (avoid ties). Rank the random numbers from the smallest (1) to the largest (30). The ranks will represent the pot numbers and assign treatment A to the first five pot numbers, B to the next five pot numbers, etc.

Example: Three treatments each replicated four times in CRD



Raw data of nitrogen content of common bean inoculated with 6 rhizobium strains. Treatments are designated with letters and nitrogen content (mg) in parenthesis.

1A (19.4)	2E (14.3)	3C (17.0)	4B (17.7)	5A (32.6)	6D (20.7)
12 B (24.8)	11C (19.4)	10A (27)	9E (11.8)	8D (21.0)	7E (14.4)
13 F (17.3)	14A (32.1)	15F (19.4)	16D (20.5)	17B (27.9)	18.C (9.1)
24D (18.6)	23F (19.1)	22B (25.2)	21E (11.6)	20C (11.9)	19D (18.8)
25C (15.8)	26E (14.2)	27B (24.3)	28F (16.9)	29A (33.0)	30F (20.8)

4.3. Analysis of variance of CRD with equal replication

1. State the model

The model for CRD expressing the relationship between the response to the treatment and the effect of other factors unaccounted for:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{where}$$

Y_{ij} = the j^{th} observation on the i^{th} treatment;

μ = General mean;

τ_i = Effect of treatment i ; and ε_{ij} = Experimental error (effect due to chance)

2. Arrange the data by treatments and calculate the treatment totals (T_i) and Grand total (G)

Nitrogen content of common bean inoculated with 6 rhizobium strains (mg)

Treatments (R-strains)	Reps					Treatment total (T _i)
3 Dok 1	19.4	32.6	27.0	32.1	33.0	144.1
3 Dok 5	17.7	24.8	27.9	25.2	24.3	119.9
3 Dok 4	17.0	19.4	9.1	11.9	15.8	73.2
3 Dok 7	20.7	21.0	20.5	18.8	18.6	99.6
3 Dok 15	14.3	14.4	11.8	11.6	14.2	66.3
Composite	17.3	19.4	19.1	16.9	20.8	93.5
Grand total (G)						596.6

3. Using Y_{ij} = j^{th} observation on the i^{th} treatment; T_i = Treatment total; $n = (r \times t)$, the total number of experimental unit (pots), calculate the correction factor and the various sum of squares

$$- \text{C.F.} = \frac{G^2}{r \times t} = \frac{(596.6)^2}{5 \times 6} = 11864.38$$

$$- \text{Total Sum of Squares (TSS)} = \sum Y_{ij}^2 - \text{C.F. (Sum of the square of all observations- C.F.)}$$

$$= [(19.4)^2 + (17.7)^2 + \dots + (20.8)^2] - 11864.38 = 12994.36 - 11864.38 = 1129.98$$

$$- \text{Treatment Sum of Squares (SST)} = \frac{\sum T_i^2}{r} - \text{C.F.}$$

$$= \frac{(144.1)^2 + (119.9)^2 + \dots + (93.5)^2}{5} - 11864.38 = 847.05$$

$$- \text{Error Sum of Squares (SSE)} = \text{Total SS} - \text{Treatment SS} = 1129.98 - 847.05 = 282.93$$

In CRD, treatment sum of squares are usually called between or among groups sum of squares while the sum of squares among individuals treated alike is called within group or error sum of squares.

4. Calculate the mean squares (MS) for treatment and error by dividing each sum of squares by the corresponding degree of freedom

$$- \text{Treatment MS} = \frac{SST}{t-1} = \frac{847.05}{(6-1)} = 169.41$$

$$- \text{Error MS} = \frac{\text{Error SS}}{t(r-1)} = \frac{282.93}{6(5-1)} = 11.79$$

5. Calculate F-value for testing significance of treatment effects

$$- \text{F-calculated} = \frac{\text{Treatment MS}}{\text{Error MS}} = \frac{169.41}{11.79} = 14.37$$

6. Obtain the tabulated F-value using treatment degree of freedom (d. f.) as numerator (n_1) and error d. f. as denominator (n_2) at 5% and 1% level of significance

$$\text{F (5, 24) at 5\%} = 2.60; \text{F (5, 24) at 1\%} = 3.90$$

7. Summarize all the values computed on the above steps in the ANOVA- table for quick assessment of results.

Source of Variation	DF	SS	MS	Computed F	Table F	
					5%	1%
Treatment (among strains)	(t-1) = 5	847.05	169.41	14.37**	2.60	3.90
Error (within strains)	t(r-1) = 24	282.93	11.79			
Total	(rt-1) = 29	1129.98				

8. Compare the calculated F- value with table F- value and decide on significance among the treatment effects using the following rules:

- If $F_{\text{calculated}} > F_{\text{table}}$ at 1% level of significance, the difference between treatments is highly significant. Put two asterisks on $F_{\text{calculated}}$
- If $F_{\text{calculated}} > F_{\text{table}}$ at 5% level of significance, but $\leq F_{\text{table}}$ at 1%, the difference between treatments is significant. Put one asterisks on $F_{\text{calculated}}$.
- If $F_{\text{calculated}} \leq F_{\text{table}}$ at 5% level of significance, the differences among treatments is non-significant. Put NS on the $F_{\text{calculated}}$ value in ANOVA table.

Note that a non-significant F test in the analysis of variance indicates the failure of the experiment to detect any difference among treatments. It does not, in any way, prove that all treatments are the same. The failure to detect treatment difference based on non- significant F-test could be the result of either a very small or nil treatment difference or a very large experimental error or both. Thus, whenever the F- test is non-significant, the researcher should examine the size of experimental error and the numerical difference among the treatment means. If both values are large, the trial may be repeated and efforts should be made to reduce experimental error so that the differences among treatments, if any can be detected. On the other hand, if both values are small, the difference among treatments is probably too small to be of any economic value and, thus, no additional trials are needed.

For the above example, the computed F value of 14.37 is larger than the tabulated F value at the 1% level of significance of 3.90. Hence, the treatment difference is said to be highly significant. In other words, chances are less than 1 in 100 that all the observed differences among the six treatment means could be due to chance. It should also be noted that such a significant F test verifies the existence of some differences among the treatments tested but does not specify the particular pair (or pairs) of

treatments that differ significantly. To obtain this information, procedures for comparing treatment means are used.

9. Compute the Coefficient of Variation (CV) and standard error (SE) of the treatment means

$$\begin{aligned}
 - \text{CV} &= \frac{\sqrt{\text{Error MS}}}{\text{Grand mean}(G/rt)} \times 100 = \frac{\sqrt{11.79}}{19.89} \times 100 = 17.3\% \\
 - \text{SE} \pm &= \sqrt{\text{MSE}/r} = \sqrt{11.79/5} = 1.53\text{mg}
 \end{aligned}$$

Coefficient of Variation indicates the degree of precision with which the treatments are compared and it is a good index of the reliability of the experiment. The smaller the CV, the more reliable the experiment is. The CV values greatly vary with the type of experiment, experimental material or the character measured (*e.g.* data on days to flowering have smaller CV than number of nodules in common bean as within treatment variation is usually small in the former parameter than the later). In field experiments CV up to **30%** are common and usually lesser CV for laboratory and greenhouse experiments are expected than field experiments.

4.4 Analysis of variance of CRD with unequal replication

Example: Twenty rats were assigned equally at random to four feed types. Unfortunately one of the rats died due to unknown reason. The data are rat body weight in g after being raised on these diets for 10 days. We would like to know whether weights of rats are the same for all four diets at 5%.

	Feed 1	Feed 2	Feed 3	Feed 4
	60.8	68.7	102.6	87.9
	57.0	67.7	102.1	84.2
	65.0	74.0	100.2	83.1
	58.6	66.3	96.5	85.7
	61.7	69.8		90.3
T _i	303.1	346.5	401.4	431.2
n _i	5	5	4	5

$$\begin{aligned}
 - \text{C.F.} &= \frac{G^2}{n} = \frac{(1482.2)^2}{19} = 115627.20 \\
 - \text{Total Sum of Squares (TSS)} &= \sum y_{ij}^2 - \text{C.F. (Sum of the square of all observations- C.F.)} \\
 &= [(60.8)^2 + (57.00)^2 + \dots + (90.3)^2] - 115627.20 = 4354.698
 \end{aligned}$$

- Treatment Sum of Squares (SST) = $\sum \frac{T_i^2}{r_i} - \text{C.F.}$

$$\frac{(303.1)^2}{5} + \frac{(346.5)^2}{5} + \frac{(401.4)^2}{4} + \frac{(431.2)^2}{5} - 115627.20 = 4226.348$$
- Error Sum of Squares (SSE) = Total Sum of Squares – Treatment Sum of Squares: 4354.698 - 4226.348 = 128.35

Source of Variation	DF	SS	MS	Computed F	F-table (1%)
Treatment (among groups)	(t-1) = 3	4226.348	1408.788	164.6**	5.42
Error (within groups)					
(total d.f. – treatment d.f.)	(18-3) = 15	128.35	8.557		
Total	(n-1) = 18	4354.698			

** Highly significant

$$\frac{\sqrt{MSE}}{\text{Grand mean}} \times 100 = \frac{\sqrt{8.557}}{1482.2/19} \times 100 = 3.75\%$$

5. Randomized Complete Block Design (RCBD)

5.1 Uses, advantages and disadvantages

It is the most frequently used experimental design in field experiments. Completely Randomized Design is appropriate when no sources of variation, other than treatment effects are known or anticipated, *i.e.* the experimental units should be similar. However, in many experiments certain experimental units if treated alike will behave differently, *e.g.* in field experiments adjacent plots are more alike in response than those some distant part. Heaviest animals in a group of the same age may show a different rate of weight gain than lighter animals.

In such situations, designs and layouts can be constructed so that the portion of variability attributable to the known source can be measured and excluded from experimental error. Thus, difference among treatment means will contain no contribution to the known source.

Randomized Complete Block Design (RCBD) can be used when the experimental units can be meaningfully grouped, the number of units in a group being equal to the number of treatments. Such a group is called a block and equals to the number of replications. Each treatment appears an equal number of times usually once, in each block and each block contains all the treatments.

Blocking (grouping) can be done based on soil heterogeneity in a fertilizer or variety trials; initial body weight, age, sex, and breed of animals; slope of the field, etc.

The primary purpose of blocking is to reduce experimental error by eliminating the contribution of known sources of variation among experimental units. By blocking variability within each block is minimized and variability among blocks is maximized.

During the course of the experiment, all units in a block must be treated as uniformly as possible. For example,

- if planting, weeding, fertilizer application, harvesting, data recording, etc, operations cannot be done in one day due to some problems, then all plots in any one block should be done at the same time.
- if different individuals have to make observations of the experimental plots, then one individual should make all the observations in a block.

This practice helps to control variation within blocks, and thus variation among blocks is mathematically removed from experimental error.

Advantages:

1. **Precision:** More precision is obtained than with CRD because grouping experimental units into blocks reduces the magnitude of experimental error.
2. **Flexibility:** Theoretically there is no restriction on the number of treatments or replications. If extra replication is desired for certain treatments, it can be applied to two or more units per block.
3. **Ease of analysis:** The statistical analysis of the data is simple. If as result of change or misshape, the data from a complete block or for certain treatments are unusable, the data may be omitted without complicating the analysis. If data from individual units (plots) are missing, they can be estimated easily so that simplicity of calculation is not lost.

Disadvantages:

The main disadvantage of RCBD is that when the number of treatments is large (>15), variation among experimental units within a block becomes large, resulting in a large error term. In such situations, other designs such as incomplete block designs should be used.

5.2 Randomization and layout

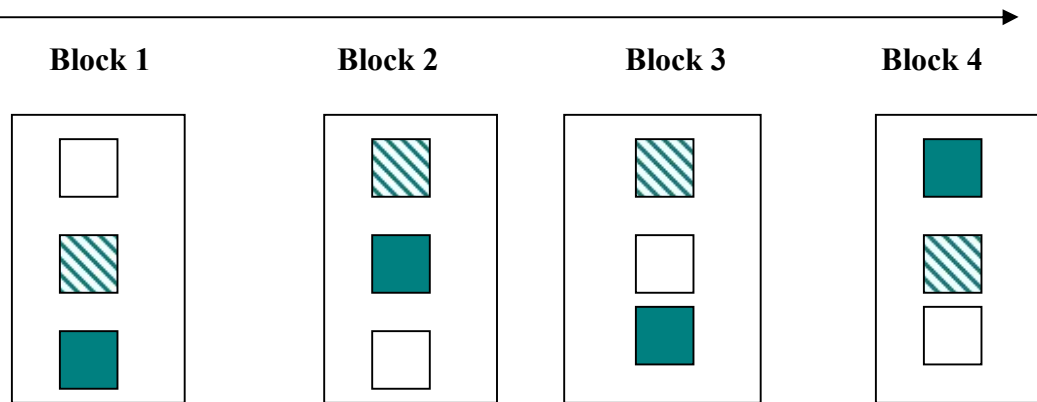
Step 1: Divide the experimental area (unit) into r -equal blocks, where r is the number of replications, following the blocking technique. Blocking should be done against the gradient such as slope, soil fertility, etc.

Step 2: Sub-divide the first block into t -equal experimental plots, where t is the number of treatments and assign t treatments at random to t -plots using any of the randomization scheme (random numbers or lottery).

Step 3: Repeat step 2 for each of the remaining blocks.

Three treatments each replicated four times

ENVIRONMENTAL GRADIENT



The major difference between CRD and RCBD is that in CRD, randomization is done without any restriction to all experimental units but in RCBD, all treatments must appear in each block and different randomization is done for each block (randomization is done within blocks).

5.3 Analysis of variance of RCBD

Step 1. State the model

The linear model for RCBD:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

where, Y_{ij} = the observation on the j^{th} block and the i^{th} treatment; μ = common mean effect; τ_i = effect of treatment i ; β_j = effect of block j ; and ε_{ij} = experiment error for treatments i in block j .

Step 2. Arrange the data by treatments and blocks and calculate treatment totals (T_i), Block (rep) totals (B_j) and Grand total (G).

Example: Oil content of linseed treated at different stages of growth with N-fertilizers.

Oil content (g) from sample of 20 g seed

Treatment (Stages of application)	Block 1	Block 2	Block 3	Block 4	Treat. Total (T_i)
Seeding	4.4	5.9	6.0	4.1	20.4
Early blooming	3.3	1.9	4.9	7.1	17.2
Half blooming	4.4	4.0	4.5	3.1	16.0
Full blooming	6.8	6.6	7.0	6.4	26.8
Ripening	6.3	4.9	5.9	7.1	24.2
Unfertilized (control)	6.4	7.3	7.7	6.7	28.1
Block total (B_j)	31.6	30.6	36.0	34.5	
Grand total					132.7

From this table we can test:

1. Does fertilizer application have effect on oil content? Comparing treatment 6 versus 1-5.

2. Which stage of application is best in increasing oil content in linseed? Comparing the difference between 1, 2, 3, 4 and 5

Step 3: Compute the correction factor (C.F.) and sum of squares using r as number of blocks, t as number of treatments, T_i as total of treatment i, and B_j as total of block j.

$$a. \text{ C.F.} = \frac{G^2}{rt} = \frac{(132.7)^2}{4 \times 6} = 733.72$$

$$b. \text{ Total Sum of Square (TSS)} = \sum Y_{ij}^2 - \text{C.F.} = (4.4)^2 + (3.3)^2 + \dots + (6.7)^2 - \text{C.F.} \\ = 788.23 - 733.72 = 54.51$$

$$c. \text{ Block Sum of Squares (SSB)} = \frac{\sum_{j=1}^r B_j^2}{t} - \text{C.F.} \\ = \frac{(31.6)^2 + (30.6)^2 + (36.0)^2 + (34.5)^2}{6} - 733.72 ; 736.86 - 733.72 = 3.14$$

$$d. \text{ Treatment Sum of Squares (SST)} = \frac{\sum_{i=1}^t T_i^2}{r} - \text{C.F.} \\ = \frac{(20.4)^2 + (17.2)^2 + (16)^2 + \dots + (28.1)^2}{4} - 733.72 ; 765.37 - 733.72 = 31.65$$

$$e. \text{ Error Sum of Squares (SSE)} = \text{Total SS} - \text{SSB} - \text{SST} \\ = 54.51 - 3.14 - 31.65 = 19.72$$

Step 4: Compute the mean squares for block, treatment and error by dividing each sum of squares by its corresponding d.f.

$$\text{Block Mean Square (MSB)} = \frac{SSB}{r-1} = \frac{3.14}{3} = 1.05$$

$$\text{Treatment Mean Square (MST)} = \frac{SST}{t-1} = \frac{31.65}{5} = 6.33$$

$$\text{Error Mean Square (MSE)} = \frac{SSE}{(r-1)(t-1)} = \frac{19.72}{15} = 1.31$$

Step 5: Compute the F-value for testing block and treatment differences.

$$\text{F-block} = \frac{MSB}{MSE} \text{ for block; } \text{F-treatment} = \frac{MST}{MSE} \text{ for treatments.}$$

$$\text{F-block} = \frac{1.05}{1.31} = 0.80 ; \text{F-treatment} = \frac{6.33}{1.31} = 4.83$$

Step 6: Read table F- and compare the computed F-value with tabulated F-value and make decision.

- F-table for comparing block effects, use block d. f. as numerator (n_1) and error d.f. as denominator (n_2); F (3, 15) at 5% = 3.29 and at 1% = 5.42.
- F-table for comparing treatment effects, use treatment d.f. as numerator and error d.f. as denominator F (5, 15) at 5% = 2.90 and at 1% = 4.56. If the calculated F-value is greater than the tabulated F-

value for treatments at 1%, it means that there is a highly significant (real) difference among treatment means.

In the above example, calculated F-value for treatments (4.83) is greater than the tabulated F-value at 1% level of significance (4.56). Thus, there is a highly significant difference among the stages of application on nitrogen content of the linseed.

Step 7: Compute the standard error of the mean and coefficient of variability.

- Standard error (SE_{\pm})

$$SE_{\pm} = \sqrt{\frac{MSE}{r}} = \sqrt{\frac{1.31}{4}} = 0.57g$$

$$CV = \frac{\sqrt{MSE}}{\text{Grand mean}} \times 100 = \frac{\sqrt{1.31}}{5.53} \times 100 = 20.7\%$$

Step 8: Summarize the results of computations in analysis of variance table for quick assessment of the result.

Analysis of variance for oil content of the linseed.

Source of variation	Degree of Freedom	Sum of Squares	Mean Squares	Computed F	Tabulated F	
					5%	1%
Block	(r-1) = 3	3.14	1.05	0.80	3.29	5.42
Treatment	(t-1) = 5	31.65	6.33	4.83**	2.90	4.56
Error	(r-1) (t-1) = 15	19.72	1.31			
Total	(rt-1) = 23	54.51				

5.4. Block efficiency

Blocking maximizes the difference among blocks and reduces the difference among plots of the same block (within blocks) as small as possible. Thus, the result of every RCBD should be examined to see whether this objective has been achieved. The procedure to measure block efficiency is:

Step 1: Determine the level of significance of block variation by computing F-value for block and test its significance.

$$F\text{-block} = \frac{MSB}{MSE} = \frac{1.05}{1.31} = 0.80$$

By comparing it with tabulated F-value at n_1 (r-1) d.f. and n_2 error d.f. (r-1) (t-1) = F (3, 15) at 5% = 3.29.

If the computed F-value is greater than the tabulated F-value, blocking is said to be effective in reducing experimental error. Also the scope of an experiment may have been increased when blocks are significantly different since the treatments have been tested over a wider range of experimental conditions.

On the other hand, if block effects are small (calculated F for block < tabulated F value), it indicates either that the experimenter was not successful in reducing error variance by grouping of individual units (blocking) or that the units were essentially homogeneous to start with.

Step 2: Determine the magnitude of the reduction in experimental error due to blocking by computing the Relative Efficiency (R.E.) as compared to CRD.

$$R.E. = \frac{(r-1)MSB + r(t-1)MSE}{(rt-1)MSE}$$

Where MSB = block mean square; MSE = error mean square; r = number of replications; t = number of treatments

$$R.E. = \frac{(4-1) \times (1.05) + 4 \times (6-1) \times 1.31}{(4 \times 6 - 1) \times 1.31}$$

$$R.E. (RCBD \text{ to CRD}) = \frac{(3 \times 1.05) + (4 \times 5 \times 1.31)}{23 \times 1.31} = \frac{3.15 + 26.2}{30.13} = \frac{29.35}{30.13} = 0.97$$

If error degree of freedom of RCBD is less than 20, the R.E. should be multiplied by the adjustment factor (k) to consider the loss in precision resulting from fewer degrees of freedom.

$$K = \frac{[(r-1)(t-1)+1][t(r-1)+3]}{[(r-1)(t-1)+3][t(r-1)+1]}$$

$$\begin{aligned} K &= \frac{[(4-1)(6-1)+1][6(4-1)+3]}{[(4-1)(6-1)+3][6(4-1)+1]} \\ &= \frac{(3 \times 5 + 1)(6 \times 3 + 3)}{(3 \times 5 + 3)(6 \times 3 + 1)} = \frac{16 \times 21}{18 \times 19} = \frac{336}{342} = 0.98 \end{aligned}$$

$$\text{Adjusted R.E.} = 0.97 \times K(0.98) = 0.95.$$

In this case, information is sacrificed in theory by using Randomized Complete Block Design, since 95 replicates in a completely randomized design give as much information as 100 blocks or replications of a Randomized Complete Block Design.

5.5. Missing Data in RCBD

Sometimes data for certain units may be missing or become unusable. For example,

- when an animal becomes sick or dies but not due to treatment
- when rodents destroy a plot in field
- when a flask breaks in laboratory
- when there is an obvious recording error

A method is available for estimating such data. Note that an estimate of a missing value does not supply additional information to the experimenter; it only facilitates the analysis of the remaining data.

Case 1: When a single value is missing

When a single value is missing in RCBD, an estimate of the missing value can be calculated as:

$$Y = \frac{rBo + tTo - Go}{(r-1)(t-1)} \text{ where;}$$

Y = estimate of the missing value

t = No. of treatments

r = No. of replications or blocks

Bo = total of observed values in block (replication) containing the missing value

To = total of observed values in treatment containing the missing value

Go = Grand total of all observed values.

The estimated value is entered in the table with the observed values and the analysis of variance is performed as usual with one d. f. being subtracted from both total and error d. f. because the estimated value makes no contribution to the error sum of squares.

When all of the missing values are on the same block or treatment the simplest solution is to consider as if the block or treatment had not been included in the experiment.

Example: In the table given below are yields (kg) of 4-varieties of maize (Al-composite, Rarree-1, Bukuri, Katumani) in 4-replications planted in RCBD on a plot size of 10m x 10 m of which one plot yield is missing. Estimate the missing value and analyze the data

Varieties	I	II	III	IV	Treatment total (Ti)
Bukuri	18.5	15.7	16.2	14.1	64.5
Katumani	11.7	—	12.9	14.4	39(To)
Rarree-1	15.4	16.6	15.5	20.3	67.8
Al-composite	16.5	18.6	12.7	15.7	63.5
Block total	62.1	50.9 (Bo)	57.3	64.5	

Go (Grand total) = 234.8

Solution

a. Estimate the missing value

$$Y = \frac{rBo + tTo - Go}{(r-1)(t-1)}$$

$$Y = \frac{[(4 \times 50.9) + (4 \times 39) - 234.8]}{(4-1)(4-1)} = 13.9$$

b. Enter the estimated value and carry out the analysis following the usual procedure:

- Corrected treatment total = 39 + 13.9 = 52.9
- Corrected block total = 50.9 + 13.9 = 64.8
- Corrected grand total = 234.8 + 13.9 = 248.7

c. Analysis of variance

1. $C.F. = \frac{(248.7)^2}{rt(4 \times 4)} = 3865.73$
2. $TSS = \sum_{i=1}^4 \sum_{j=1}^4 Y_{ij}^2 - C.F. = (18.5)^2 + (11.7)^2 + \dots + (13.9)^2 + \dots + (15.7)^2 - C.F. = 79.18$
3. Treatment SS =
$$\frac{\sum_{i=1}^4 T_i^2}{r} - C.F. = \frac{(64.5)^2 + (52.9)^2 + (67.8)^2 + (63.5)^2}{4} - 3865.73 = 31.21$$
4. Block SS =
$$\frac{\sum_{j=1}^4 B_j^2}{t} - C.F. = \frac{(62.1)^2 + (64.8)^2 + (57.3)^2 + (64.5)^2}{4} - 3865.73 = 9.02$$
5. Error SS = Total SS - Treatment SS - Block SS = $79.18 - 31.21 - 9.02 = 38.95$

d. Compute the correction factor for bias (B) for treatment sum of squares as the treatment SS is biased upwards.

$$B = \frac{[Bo - (t-1)y]^2}{t(t-1)}$$

Bo = Total of observed values in blocks (replication) containing the missing value

Y = estimated value

$$B = \frac{[50.9 - (4-1) \times 13.9]^2}{4 \times (4-1)}$$

$$= \frac{[50.9 - 41.7]^2}{12} = 7.05$$

e. Subtract the computed B value from Total SS & Treatment SS

$$\begin{aligned} \text{Adjusted Treatment SS} &= \text{Treatment SS} - B \\ &= 31.21 - 7.05 = 24.16 \end{aligned}$$

$$\begin{aligned} \text{Adjusted Total SS} &= \text{Total SS} - B \\ &= 79.18 - 7.05 = 72.13 \end{aligned}$$

f. Subtract 1 from error d. f. and total d. f. and complete the analysis of variance table.

Source of variation	Degree of freedom	Sum of squares	Mean squares	Computed F	Table F	
					5%	1%
Block	3	9.02	3.01	0.61 ^{ns}	4.07	7.59
Treatment	3	24.16	8.05	1.64 ^{ns}	4.07	7.59
Error	(t-1)(r-1)-1 = 8	38.95	4.9			
Total	rt-1-1 = 14	72.13				

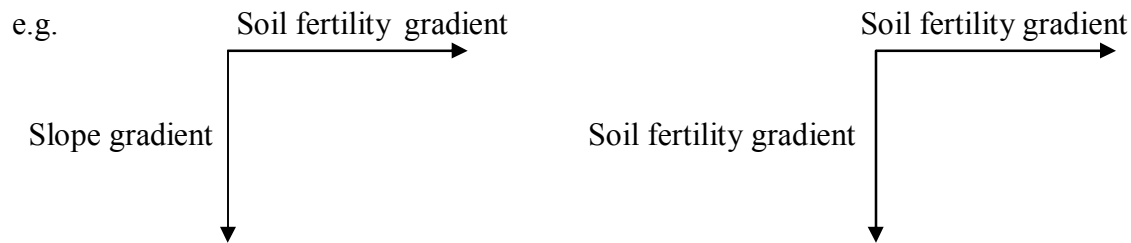
$$CV = \frac{\sqrt{MSE}}{\text{Grand mean}} \times 100 = \frac{\sqrt{4.9}}{15.65(234.8/15)} \times 100 = 14.1\%$$

6. Latin Square Design

6.1. Uses, advantages and disadvantages

The major feature of the Latin Square Design is its capacity to simultaneously handle two known sources of variation among experimental units unlike Randomized Complete Block Design (RCBD), which treats only one known source of variation.

The two directional blocking in a Latin Square Design is commonly referred as row blocking and column blocking. In Latin Square Design the number of treatments is equal to the number of replications that is why it is called Latin Square.



Advantages:

- Greater precision is obtained than Completely Randomized Design & Randomized Complete Block Design (RCBD) because it is possible to estimate variation among row blocks as well as among column blocks and remove them from the experimental error.

Disadvantages:

- As the number of treatments is equal to the number of replications, when the number of treatments is large the design becomes impractical to handle. On the other hand, when the number of treatments is small, the degree of freedom associated with the experimental error becomes too small for the error to be reliably estimated. Thus, in practice the Latin Square Design is applicable for experiments in which the number of treatments is not less than four and not more than eight.
- Randomization is relatively difficult.

6.2. Randomization and layout

Step 1: To randomize a five treatment Latin Square Design, select a sample of 5 x 5 Latin square plan from appendix of statistical books. We can also create our own basic plan and the only requirement is that each treatment must appear only once in each row and column. For our example, the basic plan can be:

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

Step 2: Randomize the row arrangement of the plan selected in step 1, following one of the randomization schemes (either using lottery method or table of random numbers).

- Select from table of random numbers, five three digit random numbers avoiding ties if any
Random numbers: 628 846 475 902 452
Rank: (3) (4) (2) (5) (1)
- Rank the selected random numbers from the lowest (1) to the highest (5)
- Use the ranks to represent the existing row number of the selected plan and the sequence to represent the row number of the new plan. For our example, the third row of the selected plan (rank 3) becomes the first row (sequence) of the new plan, the fourth becomes the second row, etc.

	1	2	3	4	5
3	C	D	A	E	B
4	D	E	B	A	C
2	B	A	E	C	D
5	E	C	D	B	A
1	A	B	C	D	E

Step 3: Randomize the column arrangement using the same procedure. Select five three digit random numbers.

Random numbers: 792 032 947 293 196
Rank: (4) (1) (5) (3) (2)

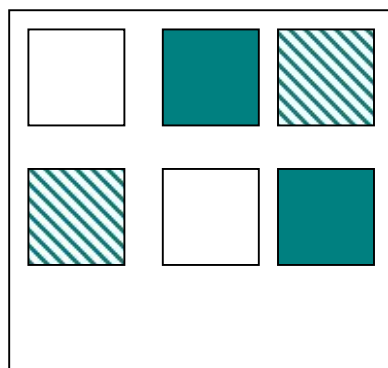
The rank will be used to represent the column number of the above plan (row arranged) in step 2. For our example, the fourth column of the plan obtained in step 2 above becomes the first column of the final plan, the first column of the plan becomes 2, etc.

Final layout:

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

Note that each treatment occurs only once in each row and column

Sample layout of three treatments each replicated three times in Latin Square Design





6.3. Analysis of variance

There are four sources of variation in Latin Square Design, two more than that of CRD and one more than that for the RCBD. The sources of variation are row, column, treatment and experimental error.

The linear model for Latin Square Design:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + \varepsilon_{ijk}$$

where,

Y_{ijk} = the observation on the i^{th} treatment, j^{th} row & k^{th} column

μ = Common mean effect

τ_i = Effect of treatment i

β_j = Effect of row j

γ_k = Effect of column k

ε_{ijk} = Experiment error (residual) effect

Example: Grain yield of three maize hybrids (A, B, and D) and a check variety, C, from an experiment with Latin Square Design.

Grain yield (t/ha)

Row number	C1	C2	C3	C4	Row total (R)
R1	1.640(B)	1.210(D)	1.425(C)	1.345(A)	5.620
R2	1.475(C)	1.15(A)	1.400(D)	1.290(B)	5.350
R3	1.670(A)	0.710(C)	1.665(B)	1.180(D)	5.225
R4	1.565(D)	1.290(B)	1.655(A)	0.660(C)	5.170
Column total(C)	6.350	4.395	6.145	4.475	
Grand total (G)					21.365

Treatment	Total (T)	Mean
A	5.855	1.464
B	5.885	1.471
C	4.270	1.068
D	5.355	1.339

STEPS OF ANALYSIS

Step 1: Arrange the raw data according to their row and column designation, with the corresponding treatments clearly specified for each observation and compute row total (R), column total (C), the grand total (G) and the treatment totals (T).

Step 2: Compute the C.F. and the various Sum of Squares

$$C.F. = \frac{G^2}{t^2} = \frac{(21.365)^2}{16} = 28.53$$

$$\text{Total SS} = \sum y^2 - C.F. = [(1.64)^2 + (1.210)^2 + \dots + (0.660)^2] - 28.53 = 1.41$$

$$\text{Row SS} = \frac{\sum R^2}{t} - C.F. = \frac{(5.62)^2 + (5.35)^2 + (5.225)^2 + (5.170)^2}{4} - 28.53 = 0.03$$

$$\text{Column SS} = \frac{\sum C^2}{t} - C.F. = \frac{(6.35)^2 + (4.395)^2 + (6.145)^2 + (4.475)^2}{4} - 28.53 = 0.83$$

$$\text{Treatment SS} = \frac{\sum T^2}{t} - C.F. = \frac{(5.855)^2 + (5.885)^2 + (4.270)^2 + (5.355)^2}{4} - 28.53 = 0.43$$

$$\text{Error SS} = \text{Total SS} - \text{Row SS} - \text{Column SS} - \text{Treatment SS} = 1.41 - 0.03 - 0.83 - 0.43 = 0.12$$

Step 3: Compute the mean squares for each source of variation by dividing the sum of squares by its corresponding degrees of freedom.

$$\text{Row MS} = \frac{\text{Row SS}}{t-1} = \frac{0.03}{3} = 0.01; \text{Column MS} = \frac{\text{Column SS}}{t-1} = \frac{0.83}{3} = 0.276$$

$$\text{Treat. MS} = \frac{\text{Treatment SS}}{t-1} = \frac{0.43}{3} = 0.143; \text{Error MS} = \frac{\text{Error SS}}{(t-1)(t-2)} = \frac{0.12}{3 \times 2} = 0.02$$

Step 4: Compute the F-value for testing the treatment effect and read table F-value as:

$$F - \text{calculated} = \frac{\text{Treatment MS}}{\text{Error MS}} = \frac{0.143}{0.02} = 7.15$$

As the computed F-value (7.15) is higher than the tabulated F-value at 5% level of significance (4.76), but lower than the tabulated F-value at the 1% level (9.78), the treatment difference is significant at the 5% level of significance.

$$\text{Compute the CV as:} = \frac{\sqrt{\text{Error MS}}}{\text{Grand mean}} \times 100 = \frac{\sqrt{0.02}}{1.335} \times 100 = 10.6\%$$

Note that although the F-test on the analysis of variance indicates significant differences among the mean yields of the 4-maize varieties tested, it does not identify the specific pairs or groups of varieties that differed significantly. For example, the F-test is not able to answer the question whether every one of the three hybrids gave significantly higher yield than that of the check variety. To answer these questions, the procedure for mean comparison should be used.

Step 5: Summarize the results of the analysis in ANOVA table

Source	D.F.	SS	MS	Computed F	Table F	
					5%	1%
Row	(t-1)=3	0.03	0.01	0.50ns	4.76	9.78
Column	(t-1)=3	0.83	0.275	13.75**	4.76	9.78
Treatment	(t-1)=3	0.43	0.142	7.15*	4.76	9.78
Error	(t-1)(t-2)=6	0.13	0.02			
Total	(t ² -1)=15	1.41				

6.4 Relative efficiency

As in RCBD, where the efficiency of one way blocking indicates the gain in precision relative to CRD, the efficiencies of both row and column blocking in a Latin Square Design indicate the gain in precision relative to either the CRD or RCBD, the procedures are:

- Compute the F-value for testing the row & column effects; and test their significance

$$F(\text{row}) = \frac{\text{Row } MS}{\text{Error } MS} = \frac{0.01}{0.02} = 0.50; \quad F(\text{column}) = \frac{0.275}{0.02} = 13.75^{**}$$

- Compute the relative efficiency of LS design relative to CRD & RCBD

The relative efficiency of LS design as compared to CRD

$$\text{R.E. (CRD)} = \frac{MS_{\text{row}} + MS_{\text{column}} + (t-1) \times MS_{\text{error}}}{(t+1) \times (MS_{\text{error}})}$$

$$= \frac{0.01 + 0.275 + (4-1) \times 0.02}{(4+1) \times 0.02} = \frac{0.345}{0.100} = 3.45$$

This indicates that the use of Latin Square Design in the present example is estimated to increase the experimental precision by 245% as compared to CRD. This result implies that if the CRD is used an estimated 2.45 times more replication would have been required to detect the treatment difference of the same magnitude as that detected with the Latin Square Design.

- The R.E. of LS design as compared to RCBD can be computed in two ways:

- When row is used as blocking factor

$$\text{R.E. (row)} = \frac{MS_{\text{row}} + (t-1) \times MS_{\text{error}}}{t \times (MS_{\text{error}})} = \frac{0.01 + (4-1) \times 0.02}{4 \times 0.02} = \frac{0.07}{0.08} = 0.875$$

- When column used as blocking factor

$$\text{R.E. (column)} = \frac{MS_{\text{column}} + (t-1) \times MS_{\text{error}}}{t \times (MS_{\text{error}})} = \frac{0.275 + (4-1) \times 0.02}{4 \times 0.02} = \frac{0.335}{0.08} = 4.19$$

When the error d. f. in the Latin Square analysis of variance is < 20, the R.E. value should be multiplied by the adjustment factor (K) defined as:

$$K = \frac{[(t-1)(t-2)+1][(t-1)^2+3]}{[(t-1)(t-2)+3][(t-1)^2+1]} = \frac{[(4-1)(4-2)+1][(9+3)]}{[(4-1)(4-2)+3][10]} = \frac{7 \times 12}{9 \times 10} = \frac{84}{90} = 0.93$$

The adjusted R.E. values are computed as:

$$\text{R.E. (row)} = 0.875 \times 0.93 = 0.81$$

$$\text{R.E. (column)} = 4.19 \times 0.93 = 3.90$$

The results indicate that the additional column blocking made possible by the use of Latin Square Design is estimated to have increased the experimental precision over that of RCBD by 290%, whereas the additional row-blocking in the LS design did not increase precision over the RCBD with column as blocks. Hence, for the above trial, a RCBD with column as blocks would have been as efficient as a Latin Square Design.

6.5. Missing data

The formula for a single missing observation

$$Y = \frac{t(R_o + C_o + T_o) - 2G_o}{(t-1)(t-2)}$$

Where R_o , C_o , and T_o are the totals of the observed values for the row, column, and treatment containing the missing value, respectively, and G_o is the grand total of the observed values, and t is the number of treatments.

The analysis of variance is performed in the usual manner after entering the estimated value with one degree of freedom being subtracted from total and error degrees of freedom for each missing value.

As in the case of RCBD, the treatment sum of squares is biased upward by:

$$\text{Bias (B)} = \frac{[Go - Ro - Co - (t-1)T_o]^2}{[(t-1)(t-2)]^2}$$

Where G_o , R_o , C_o , T_o and t are as described above. Then B is subtracted from treatment SS & total SS.

Example: Yield (kg) of five rice varieties tested in Latin Square Design from plot size of 100 m².

E(12)	C (-)	B(11)	A (10)	D (8)
A (7)	D (8)	C(8)	B(7)	E (13)
C (12)	B (6)	D(7)	E (11)	A (9)
B (4)	E (10)	A (7)	D (6)	C (8)
D (5)	A (8)	E (15)	C(9)	B (5)

Estimate the missing value, complete the analysis of variance and compare the variety C with D, and A with E at 5% level of significance using LSD test.

a. Estimate the missing value

$$Y = \frac{t(R_o + C_o + T_o) - 2G_o}{(t-1)(t-2)}$$

$$Y = \frac{5(41 + 32 + 37) - 2 \times 206}{(5-1)(5-2)} = 11.5$$

b. Enter the estimated value and carry out the analysis following the usual procedure:

- Corrected row total = 41.0 + 11.5 = 52.5
- Corrected column total = 32.0 + 11.5 = 43.5
- Corrected treatment total = 37 + 11.5 = 48.5
- Corrected grand total = 206.0 + 11.5 = 217.50

c. Compute the C.F. and the various Sum of Squares

$$C.F. = \frac{G^2}{t^2} = \frac{(217.5)^2}{25} = 1892.25$$

$$\text{Total SS} = \sum y^2 - C.F. = [(12.0)^2 + (11.5)^2 + \dots + (5.0)^2] - 1892.25 = 180.0$$

$$\text{Row SS} = \sum_t R^2 - C.F. = \frac{(52.5)^2 + (43.0)^2 + \dots + (42)^2}{5} - 1892.25 = 31.6$$

$$\text{Column SS} = \sum_t C^2 - C.F. = \frac{(40.0)^2 + (43.5)^2 + \dots + (43.0)^2}{5} - 1892.25 = 6.60$$

$$\text{Treatment SS} = \sum_t T^2 - C.F. = \frac{(41.0)^2 + (33.0)^2 + (48.5)^2 + \dots + (61.0)^2}{5} - 1892.25 = 107.6$$

$$\text{Error SS} = \text{Total SS} - \text{Row SS} - \text{Column SS} - \text{Treatment SS} = 180.0 - 31.6 - 6.6 - 107.6 = 34.2$$

d. Compute the correction factor for bias (B) for treatment sum of squares as the treatment SS is biased upwards.

$$\text{Bias (B)} = \frac{[Go - Ro - Co - (t-1)T_o]^2}{[(t-1)(t-2)]^2} = \frac{[206 - 41 - 32 - (5-1)37]^2}{[(5-1)(5-2)]^2} = 1.56$$

e. Subtract the computed B value from Total SS & Treatment SS

$$\text{Adjusted Treatment SS} = \text{Treatment SS} - B = 107.6 - 1.56 = 106.04$$

$$\text{Adjusted Total SS} = \text{Total SS} - B = 180.0 - 1.56 = 178.44$$

f. Subtract 1 from error d. f. and total d. f. and complete the analysis of variance table.

Source	D.F.	SS	MS	Computed F	Table F	
					5%	1%
Row	(5-1)=4	31.6	7.90	2.54	3.36	5.67
Column	(5-1)=4	6.6	1.65	0.53	3.36	5.67
Treatment	(5-1)=4	106.04	26.51	8.52**	3.36	5.67
Error	(5-1)(5-2)-1 = 11	34.2	3.11			
Total	(t ² -1)-1 = 23	178.44				

$$CV = \frac{\sqrt{MSE}}{\text{Grand mean}} \times 100 = \frac{\sqrt{3.11}}{(206/24)} \times 100 = \frac{\sqrt{3.11}}{8.58} \times 100 = 20.5\%$$

To compare treatments C & D, $LSD_{5\%} = t_{0.025(11)} \times s\bar{d}$ where $s\bar{d} = \sqrt{MSE \left[\frac{2}{t} + \frac{1}{(t-1)(t-2)} \right]}$;

$$s\bar{d} = \sqrt{3.11 \times \left[\frac{2}{5} + \frac{1}{(5-1)(5-2)} \right]} = 1.226$$

$$LSD_{5\%} = 2.201 \times 1.226 = 2.70 \text{ kg}$$

Difference between the treatment means of C & D $(37/4 - 34/5) = 9.25 - 6.8 = 2.45$. Since the difference is less than LSD value, there is no significant difference between treatments C & D.

To compare the treatments A & E both with equal replication (without missing value)

$$LSD_{5\%} = t_{0.025(11)} \times s\bar{d}, \text{ where } s\bar{d} = \sqrt{\frac{2MSE}{r}} = \sqrt{\frac{2 \times 3.11}{5}} = 1.115$$

$$LSD_{5\%} = 2.201 \times 1.115 = 2.45 \text{ kg}$$

Difference between the treatment means of A & E $(61/5 - 41/5) = 12.2 - 8.2 = 4.00$. Since the difference is greater than LSD value, there is significant difference between treatments A & E.

7. INCOMPLETE BLOCK DESIGNS

Theoretically, the complete block designs (where each block contains all the treatments) such as Randomized Complete Block and Latin Square are applicable to experiments with any number of treatments. However, these complete block designs become less efficient as the number of treatments increases, mainly because block size increases proportionally with the number of treatments which in turn increases experimental error.

An alternative set of designs for single factor experiments having a large number of treatments are the incomplete block designs. For example, plant breeders are often interested in making comparisons among a large number of selections in a single trial. For such trials, we use incomplete block designs. As the name implies, the experimental units in these designs are grouped into blocks which are smaller than a complete replication of the treatments. However, the improved precision with the use of an incomplete block designs (where the blocks do not contain all the treatments) is achieved with some costs. The major ones are:

- inflexible number of treatments or replications or both.
- unequal degree of precision in the comparison of treatment means.
- complex data analysis.

Although there is no concrete rule as to how large the number of treatments should be before the use of an incomplete block design, the following points may be helpful:

- a. Variability in the experimental material: The advantage of an incomplete block design over complete block designs is enhanced by an increased variability in the experimental material. In general,

whenever the block size in Randomized Complete Block Design is too large to maintain reasonable level of uniformity among experimental units within the same block, the use of an incomplete block design should be seriously considered.

- b. Computing facilities and services: Data analysis of an incomplete block design is more complex than that for a complete block design. Thus, the use of an incomplete block design should be considered only as the last measure.

7.1. Lattice Designs

The lattice designs are the most commonly used incomplete block designs in agricultural experiments. There is sufficient flexibility in the design to make its applications simpler than most of the other incomplete block designs. There are two kinds of lattices: balanced lattice and partially balanced lattice designs.

Field arrangement & randomization:

- Blocks in the same replication should be made as nearly alike as possible
- Randomize the order of blocks within replication using a separate randomization in each replication.
- Randomize treatment code numbers separately in each block.

7.1 Balanced lattices

The balanced lattice design is characterized by the following basic features:

- a. Each treatment occurs together in the same block with every other treatment once. As result the statistical analysis is relatively simple and each pair of treatments is compared with the same precision. For example, consider the following design for four treatments on blocks of size two.

<u>Block</u>	<u>Rep I</u>	<u>Rep II</u>	<u>Rep III</u>
(1)	A B	(1) A C	(1) A D
(2)	C D	(2) B D	(2) B C

Note that treatment A occurs with treatment B, with treatment C and with treatment D only once.

- b. The number of treatments (t) must be perfect square such as 16, 25, 36, 49, 64, etc.
- c. The block size (k) is equal to the square root of the number of treatments, *i.e.* $k = \sqrt{t}$.
- d. The number of replications (r) is one more than the block size, *i.e.* $r = k + 1$. That is, the number of replications required is 6 for 25 treatments, 7 for 36 treatments, and so on.

As balanced lattices require large number of replications, they are not commonly used.

7.2 Partially balanced lattices

The partially balanced lattice design is more or less similar to the balanced lattice design, but it allows for a more flexible choice of the number of replications. The partially balanced lattice design requires that the number of treatments must be a perfect square and the block size is equal to the square root of the number of treatments. However, any number of replications can be used in partially balanced lattice design. The partially balanced lattice design with two replications is called simple lattice, with three replications is

Replication	Block	% kill				Block total (B)		M	Cb
1	1	75(15)	57(16)	71(13)	77(14)	280	789	-51	
	2	78(12)	66(11)	68(10)	45(9)	257	716	-55	
	3	40(6)	64(5)	49(8)	42(7)	195	608	23	
	4	59(3)	53(1)	46(2)	54(4)	212	642	6	
Rep total (R1)						944		-77	
2	1	53(16)	66(4)	57(12)	47(8)	223	663	-6	
	2	80(14)	48(6)	73(10)	52(2)	253	700	-59	
	3	36(7)	63(11)	67(15)	47(3)	213	676	37	
	4	68(13)	60(1)	50(9)	76(5)	254	716	-46	
Rep total (R2)						943		-74	
3	1	66(15)	46(2)	58(12)	69(5)	239	754	37	
	2	46(4)	40(7)	59(13)	55(10)	200	678	78	
	3	43(9)	55(3)	50(8)	68(14)	216	670	22	
	4	60(11)	58(1)	48(16)	47(6)	213	653	14	
Re total (R3)						868		151	
Grand Total (G)							2755		0

Analysis of variance

Step 1: Calculate the block total (B), the replication total (R) and Grand Total (G) as shown above.

Step 2: Calculate the treatment totals (T) by summing the values of each treatment from the three replications.

Treat. No.	Treatment total (T)	Sum Cb	A×Cb (0.0813×Cb)	Adjusted treatment total (T') (T+Acb)	Adjusted treatment mean (Ti/3)
1	171	-26	-2.11	168.89	56.29
2	144	-16	-1.30	142.70	47.57
3	161	65	5.28	166.28	55.43
4	166	78	6.34	172.34	57.45
5	209	14	1.14	210.14	70.05
6	135	-22	-1.79	133.21	44.40
7	118	138	11.22	129.22	43.07
8	146	39	3.17	149.17	49.72
9	138	-79	-6.42	131.58	43.86
10	196	-36	-2.93	193.07	64.36
11	189	-4	-0.32	188.67	62.89
12	193	-24	-1.95	191.05	63.68
13	198	-19	-1.54	196.45	65.48
14	225	-88	-7.15	217.85	72.61
15	208	23	1.87	209.87	69.96
16	158	-43	-3.49	154.50	51.50

Step 3: Using r as number of replications and k as block size, compute the total sum of squares, replication sum of squares, treatment (unadjusted) sum of squares as:

- Correction Factor (C. F.) = $\frac{G^2}{rk^2} = \frac{(2755)^2}{3 \times 16} = 158125.5$
- Total Sum of Squares = $(75)^2 + (57)^2 + \dots + (47)^2 - \text{C.F.} = 164233 - 158125.5 = 6107.5$
- Replication Sum of Squares = $\frac{\sum R^2}{k^2} - \text{C. F.} = \frac{(944)^2 + (943)^2 + (868)^2}{16} - 158125.5 = 158363.06 - 158125.5 = 237.6$
- Treatment (unadjusted) Sum of Squares = $\frac{\sum T^2}{r} - \text{C. F.}$
 $= \frac{(171)^2 + (144)^2 + \dots + (158)^2}{3} - 158125.5 = 163029 - 158125.5 = 4903.5$

Step 4: For each block, calculate block adjustment factor (Cb) as:

$Cb = M - rB$ where M is the sum of treatment totals for all treatments appearing in that particular block, B is the block total and r is the number of replications. For example, block 2 of replication 2 contains treatments 14, 6, 10, and 2. Hence, the M value for block 2 of replication 2 is: $M = T_{14} + T_6 + T_{10} + T_2 = 225 + 135 + 196 + 144 = 700$ and the corresponding Cb value is: $Cb = 700 - (3 \times 253) = -59$. The Cb

values for the blocks are presented in the above table. Note that the sum of Cb values over all replications should add to zero (*i. e.* $-77 + -74 + 151 = 0$).

Step 5: Calculate the Block (adju.) Sum of Squares as:

$$\begin{aligned} \text{Block (adj.) SS} &= \frac{\sum Cb^2}{k \times r \times (r-1)} - \frac{\sum Rc_b^2}{(k^2)(r)(r-1)} \\ &= \frac{(-51)^2 + (-55)^2 + \dots + (14)^2}{4 \times 3 \times 2} - \frac{(-77)^2 + (-74)^2 + (151)^2}{16 \times 3 \times 2} = 888.58 - 356.31 = 532.27 \end{aligned}$$

Step 6: Calculate the intra-block error sum of squares as:

$$\begin{aligned} \text{Intra-block error SS} &= \text{Total SS} - \text{Rep. SS} - \text{Treatment (unadjusted) SS} - \text{Block (adjusted) SS} \\ &= 6107.5 - 237.6 - 4903.5 - 532.27 = 434.13 \end{aligned}$$

Step 7: Calculate the intra-block error mean square (MS) and block (adj.) mean square (MS) as:

$$\begin{aligned} \text{Intra-block error MS} &= \frac{\text{Intra-block error SS}}{(k-1)(rk-k-1)} = \frac{434.13}{21} = 20.67 \\ \text{Block (adjusted) MS} &= \frac{\text{Block (adjusted) SS}}{(r)(k-1)} = \frac{532.27}{9} = 59.14 \end{aligned}$$

Note that if the adjusted block mean square is less than intra-block error mean square, no further adjustment is done for treatment. In this case, the F-test for significance of treatment effect is made in the usual manner as the ratio of treatment (unadjusted) mean square and intra-block error mean square and steps 8 to 13 can be ignored. For our example, the MSB value of 59.14 is greater than the MSE value of 20.67, thus, the adjustment factor is computed.

Step 8: Calculate adjustment factor A. For a triple lattice design, the formula is:

$$A = \frac{(Eb - Ee)}{k(r-1)(Eb)} = \frac{(59.14 - 20.67)}{4 \times 2 \times 59.14} = 0.0813$$

where E_b is the block (adju.) mean square and E_e is the intra-block error mean square, and k is block size.

Step 9: For each treatment, calculate the adjusted treatment total (T') as:

$T' = T + A \sum Cb$ where the summation runs over all blocks in which that particular treatment appears. For example, the adjusted treatment totals for treatment number 1 and 2 are computed as:

$$\begin{aligned} T'1 &= 171 + 0.0813(6 + -46 + 14) = 168.89 \\ T'2 &= 144 + 0.0813(6 + -59 + 37) = 142.70 \\ &\cdot \\ &\cdot \\ &\cdot \\ &\text{etc.} \end{aligned}$$

The adjusted treatment totals (T') and their respective means (adjusted treatment total divided by the number of replications (3)) are presented along with the unadjusted treatment totals (T) in the table above.

Step 10: Compute the adjusted Treatment Sum of Squares as:

$$\text{Treatment (unadjusted) SS} - [Ak(r-1) \frac{(r)(SSB_{un})}{(r-1)(1+kA)} - SSB]$$

$$- SSB_{un} (\text{unadjusted block sum of squares}) = \frac{\sum B^2}{k} - \text{C. F.} - \text{SSR; where B is Block total, and SSR is}$$

Replication Sum of Squares.

$$= \frac{(280)^2 + (257)^2 + \dots + (213)^2}{4} - 158125.5 - 237.6$$

$$= 160046.75 - 158125.5 - 237.6 = 1683.65$$

Thus, the adjusted treatment sum of squares

$$= 4903.5 - [(0.0813 \times 4 \times 2) \frac{(3) \times (1683.65)}{(2) \times (1 + 4 \times 0.0813)} - 532.27] = 4010.20$$

Step 11: Compute the treatment (adjusted) mean square as:

$$- \text{Treatment (adj.) MS} = \frac{\text{Treatment (adjusted) SS}}{k^2 - 1} = \frac{4010.2}{15} = 267.35$$

Step 12: Compute the F-value for testing the significance of treatment difference and compare the computed F-value with the tabulated F-value with $k^2 - 1 = 15$ degree of freedom as numerator and $(k - 1)(rk - k - 1) = 21$ as denominator.

$$- \text{Computed F} = \frac{\text{Treatment (adjus) MS}}{\text{Intra - block error MS}} = \frac{267.35}{20.67} = 12.93$$

Step 13: Compare the computed F-value with the table F-value

$$F_{0.05}(15, 21) = 2.18$$

$$F_{0.01}(15, 21) = 3.03$$

Since the computed F-value of 12.93 is greater than the tabulated F-value at 1% level of significance (3.03), the differences among the herbicide means are highly significant.

Step 14: Enter all values computed in the analysis of variance table

Source of Variation	Degrees of freedom	Sum of squares	Mean squares	Computed F	Tabulated F	
					5%	1%
Replication (r-1)	2	237.6	118.8			
Block (adj.) r (k-1)	9	532.27	59.14			
Herbicide (unadj.) k ² -1	15	4903.5	326.9			
Intra-block error (k-1) (rk-k-1)	21	434.13	20.67			
Herbicide (adj.) k ² -1	(15)	(4010.2)	267.35	12.93	2.18	3.03
Total rk ² - 1	47	6107.5				

Step 15: Compute the corresponding Coefficient of Variation (CV) as:

$$CV = \frac{\sqrt{\text{Intra-block error mean square}}}{\text{Adjusted Grand mean}} \times 100 = \frac{\sqrt{20.67}}{57.4} \times 100 = 7.9\%$$

Note that the grand mean is same for adjusted and unadjusted treatment total.

Step 16: Compute the gain in precision of the triple lattice relative to that of Randomized Complete Block Design as:

- % Relative precision = $\left[\frac{(SSB + SSE)}{(k^2 - 1)(r - 1)} \right] / Ee' \times 100$; where SSB is Block (adj.) SS, SSE is intra-block error SS, r is number of replication, and k is block size.
- Ee' (effective error mean square) = $(1 + \frac{rkA}{k+1}) \times Ee$, where r is the number of replications, k is block size, A is adjustment factor, Ee is intra-block error mean square.
- Erb = $\frac{(532.27 + 434.13)}{(4^2 - 1)(3 - 1)} / (1 + \frac{3 \times 4 \times 0.0813}{(4 + 1)}) \times 20.67$

Thus, the relative precision = $(32.2/24.7) \times 100 = 130.4$

This indicates that the precision of this experiment was increased by about 30.4% by using the triple lattice instead of Randomized Complete Block Design.

To compare between adjusted treatment mean differences

$$\bar{sd} = \sqrt{\frac{2Ee}{r} + (1 + rA)} ; \text{ where Ee is intra-block error mean square; r is number of replications and A is adjustment factor}$$

$$\text{LSD}_{1\%} = t_{0.005}(21) \sqrt{\frac{2 \times 20.67}{3} + [1 + (3 \times 0.0813)]} = 2.831 \times 3.88 = 10.97\%$$

7.3. Augmented Block Design

Augmented Block Design is used:

- when there are more number of entries/genotypes
- when there is no enough seed for test entries
- when there is no sufficient fund and land resources for replicated trials

It is not powerful design, but it is used for preliminary screening of genotypes/entries. There is no assumption of block homogeneity. Any new material is not replicated; it appears only once in the experiment while check varieties/entries occur as the number of the blocks.

- Error d. f. = (b-1) (c-1) where b = No. of blocks and c = Number of checks.

Thus, the minimum number of checks is 2 because error d. f. should be 12 or more for valid comparison. If number of checks is 1, error d.f. becomes zero which is not valid.

Randomization and Layout

Divide the experimental field into blocks. The block size may not be equal, *i.e.* some blocks may contain 10 entries while the others may contain 12 entries. Suppose we have 50 test cultures/progenies and 4 checks, we need at least 5 blocks since the error d. f. (c - 1) (b - 1) should be ≥ 12 . The higher the number of blocks, the higher the precision. If we have 5 equal blocks, block size will be 14 (10 test culture + 4 checks), but block size may vary.

Randomization

A. Conveniency

For identification purpose, sometimes the checks are assigned at the start, end or in certain intervals in the block.

Block ₁ = P₁ P₂ P₃ P₄ P₁₀ A B C D

Block ₂ = P₁₁ P₁₂ P₂₀ B C D A

Block ₃ = P₂₁ P₂₂ P₃₀ C D A B

.

.

.

etc.

where P₁, P₂, P₃₀ are test entries and A, B, C, D are checks.

B. Randomly allocate the checks in a block.

b₁ = P₁ P₂ A P₃ B P₄ D P₅ C P₆ P₇ P₁₀

b₂ = D P₁ P₂ C P₃ P₅ A B

etc.

One of the blocks can be 10 (test culture) + 4 (checks), while the other 12 (test cultures) + 4 (checks). Missing test entries (P_i) do not create problem in analysis as the analysis can be done with existing genotypes. But when checks are missing, the analysis becomes complicated.

Assume we want to test 16 new rice genotypes (test cultures) = P_1, P_2, \dots, P_{16} in block size of 4 with 4 checks to screen for early maturity.

b (number of blocks) = 4

c (number of checks) = 4 (A B C D).

Block size = 4 test cultures + 4 checks = 8

Days to maturity of rice genotypes

Block 1

P_1	A	P_2	B	P_3	C	P_4	D
120	83	100	77	90	70	85	65

Block 2

P_5	B	P_6	C	P_7	A	P_8	D
88	76	130	71	105	84	110	64

Block 3

P_9	D	P_{10}	A	P_{11}	B	P_{12}	C
102	63	140	86	135	78	138	69

Block 4

P_{13}	A	P_{14}	D	P_{15}	C	P_{16}	B
84	82	90	63	95	68	103	75

Steps of Analysis of Variance

Preliminary steps

1. Calculate block total for each block.

$$b_1 = 690 (120 + 83 + \dots + 65); b_2 = 728; b_3 = 811; b_4 = 660$$

2. Calculate total of progenies/test cultures in a particular block.

$$P \text{ block}_1 = 395; P \text{ block}_2 = 433; P \text{ block}_3 = 515; P \text{ block}_4 = 372$$

3. Construct check by block two-way table and calculate check total, check mean, check effect, sum of check totals & total of check means.

Check/ block	b_1	b_2	b_3	b_4	Check total	Check mean	Check effect (check mean- adjusted grand mean)	Check total x Check effect
A	83	84	86	82	335	83.75	-16.67	-5584.45
B	77	76	78	75	306	76.50	-23.92	-7319.52
C	70	71	69	68	278	69.50	-30.92	-8595.76

D	65	64	63	63	255	63.75	-36.67	-9350.85
Sum					1174			-30850.58
Total of check means						293.5		

Block s	n _i	Block total (B _j)	Total of test culture in a block (B _{ti})	No. of test cultures in a block (T _i)	Block effect(B _e)	T _i x B _e	Block effect x Block total
B ₁	8	690	395	4	0.375	1.5	258.75
B ₂	8	728	433	4	0.375	1.5	273.00
B ₃	8	811	515	4	0.625	2.5	506.87
B ₄	8	660	372	4	-1.375	-5.5	-907.5
Σ	32	2889			0	0	131.12

$$- b_1 = \frac{1}{4} \times (690 - 293.5 - 395) = 0.375$$

$$- b_2 = \frac{1}{4} \times (728 - 293.5 - 433) = 0.375$$

$$- b_3 = \frac{1}{4} \times (811 - 293.5 - 515) = 0.625$$

$$- b_4 = \frac{1}{4} \times (660 - 293.5 - 372) = -1.375$$

Where n_i is number of entries (test culture + checks) in each block = 4 + 4 = 8, Σn_i = N = 32.

3.1. Estimation of block effect.

$$- b_i = \frac{1}{\text{no. of checks}(C)} \times (\text{Total of the } i^{\text{th}} \text{ block} - \text{total of all check means} - \text{total of all progenies/test cultures in the block}); \sum b_i = 0$$

3.2. Calculate adjusted grand mean as:

$$\frac{1}{\text{No. of checks} + \text{No. progenies / test culture}} \times [\text{Grand total} - (b - 1) (\text{total of all check mean})] - \sum \text{no. of test cultures in a block} \times \text{corresponding block effect}]$$

Grand total = ΣB_i (sum of block total) or sum of all observations = 2889.0

$$\text{Adjusted grand mean} = \frac{1}{20} \times [2889 - (4 - 1) (293.5) - 0] = \frac{2889 - 880.5}{20} = 100.42$$

Step 3.3. Estimate check effect (C_i):

C_i = Check mean - adjusted grand mean

$$C_1 = 83.75 - 100.42 = -16.67; C_2 = 76.5 - 100.42 = -23.92$$

$$C_3 = 69.50 - 100.42 = -30.92; C_4 = 63.75 - 100.42 = -36.67$$

There are as many checks effects as the number of checks (4 in this case).

Step 3.4. Adjusted progeny value per i^{th} progeny (P_i) as:

Observed (unadjusted) progeny value - effect of block in which the i^{th} progeny is occurring
 P_1 (adjusted) = P_1 - (block effect) = $120 - (+ 0.375) = 119.62$, etc.

Progeny/ test culture no. (P_i)	Observed progeny value (P_o)	Block effect (b_i)	Adjusted progeny value ($P_o - b_i$)	Progeny effect (Adjusted progeny value-Adjusted grand mean)	Observed progeny value \times progeny effect
1	120	0.375	119.625	19.205	2304.6
2	100	0.375	99.625	-0.795	-79.5
3	90	0.375	89.625	-10.795	-971.6
4	85	0.375	84.625	-15.795	-1343
5	88	0.375	87.625	-12.795	-1126
6	130	0.375	129.625	29.205	3796.7
7	105	0.375	104.625	4.205	441.53
8	110	0.375	109.625	9.205	1012.6
9	102	0.625	101.375	0.955	97.41
10	140	0.625	139.375	38.955	5453.7
11	135	0.625	134.375	33.955	4583.9
12	138	0.625	137.375	36.955	5099.8
13	84	-1.375	85.375	-15.045	-1264
14	90	-1.375	91.375	-9.045	-814.1
15	95	-1.375	96.375	-4.045	-384.3
16	103	-1.375	104.375	3.955	407.37

		5			
Sum	1715				17216

Step 3.5. Estimate progeny effect as: Adjusted progeny value - adjusted grand mean;
For example, progeny effect for progeny 1 = 119.62-100.42 = 19.20; etc.

Analysis of variance

$$1. \text{ Correction Factor (C.F.)} = \frac{G^2}{N} = \frac{(\text{Grand total})^2}{\text{No. of observations}} = \frac{(2889)^2}{32} = 260822.53$$

$$2. \text{ Total SS} = \sum Y^2 - \text{C.F.} = \text{Sum of the squares of all observations} - \text{C.F.} = (120)^2 + (83)^2 + \dots + (75)^2 - 260822.53 = 276621 - 260822.53 = 15798.47$$

$$3. \text{ Crude block SS} = \sum \frac{B_i^2}{n_i} = \frac{(690)^2}{8} + \frac{(728)^2}{8} + \frac{(811)^2}{8} + \frac{(660)^2}{8} = 262425.62$$

$$4. \text{ True block SS: Crude block SS} - \text{C.F.} = 262425.62 - 260822.53 = 1603.09$$

$$5. \text{ Adjusted SS Due to entries (C + P)} = (\text{Adjusted grand mean} \times \text{Observed grand total}) + \\ [(\sum \text{Block effect} \times \text{corresponding block total}) + (\sum \text{Check effect} \times \text{corresponding check total}) + \\ (\sum \text{Progeny effect} \times \text{corresponding observed progeny value}) - (\text{Crude block sum squares})]$$

$$= (100.42 \times 2889) + [(131.12) + (-30850.58) + 17216 - 262425.62] = 14184.3$$

6. Unadjusted SS due to entries

$$= \frac{\sum (\text{each check total})^2}{b} + \text{Crude SS due to progenies} - \text{C.F.}$$

$$= \frac{A^2 + B^2 + C^2 + D^2}{b} + (P_1^2 + P_2^2 + \dots + P_{16}^2) - \text{C.F.}$$

b = No. of blocks (4)

$$\frac{(335)^2 + (306)^2 + (278)^2 + (255)^2}{4} + (120)^2 + (100)^2 + \dots + (103)^2 - 260822.53$$

$$= 87042.5 + 189557 - 260822.53 = 15776.97$$

7. Partition SS due to entries (C + P) to components:

$$7.1. \text{ SS due to checks} = \frac{\sum C_i^2}{b} - \frac{(\sum C_i)^2}{bc} \\ = \frac{(335)^2 + (306)^2 + \dots + (255)^2}{4} - \frac{(335 + 306 + 278 + 255)^2}{4 \times 4} \\ = \frac{(348170)}{4} - \frac{(1174)^2}{16} = 87042.5 - 86142.25 = 900.25$$

b = No. of blocks; c = No. of checks (4)

7.2. SS due to test cultures/progenies

$$= \sum P_i^2 - \frac{(\sum P_i)^2}{P} = (120)^2 + (100)^2 + \dots + (103)^2 - \frac{(1715)^2}{16}$$

$$= 189557 - 183826.56 = 5730.44$$

P = No. of progenies/test cultures = 16

7.3. SS due to checks \times test cultures/progenies = Unadjusted SS due to entries (C + P) - SS due to checks - SS due to test culture/progenies

$$= 15776.97 - 900.25 - 5730.44 = 9146.28$$

8. SS due to error: Total SS – True blocks SS - SS due to entries (adjusted)

$$= 15798.47 - 1603.09 - 14184.3 = 11.08$$

9. Summarize the results of analysis in ANOVA table

Source	D.F.	SS	MS	F-cal.	F-Table	
					5%	1%
Block (b-1)	3	1603.09	534.36			
Adjusted entries (C+P-1)	19	14184.3	746.54			
Unadjusted entries(C+P-1)	19	15776.97	830.37			
. Checks (C-1)	3	900.25	300.08	243.97**	3.86	6.99
. Test culture (P-1)	15	5730.44	382.03	310.59**	3.01	4.96
. Test culture \times check	1	9146.28	9146.28	7436.00**	5.12	10.56
Error (b-1) (C-1)	9	11.08	1.23			
Total (N – 1)	31	15798.47				

10. Mean Comparison

i. To compare any two check means at 5% level of significance

$$\text{LSD } 5\% = t_{0.025}(9) \times s_d = 2.262 \times \sqrt{\frac{2 \times \text{MSE}}{b}} = 2.262 \times \sqrt{\frac{2 \times 1.23}{4}} = 1.77 \text{ days}$$

ii. To compare two progenies/test materials occurring in the same block at 5% level of significance

$$t_{0.025}(9) \times s_d = 2.262 \times \sqrt{2 \times \text{MSE}} = \text{LSD}_{5\%} = 2.262 \times \sqrt{2 \times 1.23} = 3.55 \text{ days}$$

iii. To compare two test cultures (progenies) occurring at different block at 5% level of significance

$$\begin{aligned} \text{LSD } 5\% &= t_{0.025}(9) \times \sqrt{2 \times \text{MSE} \left(1 + \frac{1}{c}\right)}; c = \text{No. of checks} \\ &= 2.262 \times \sqrt{2 \times 1.23 \left(1 + \frac{1}{4}\right)} = 3.97 \text{ days} \end{aligned}$$

iv. To compare a progeny/test culture with any check at 5% level of significance

$$\text{LSD } 5\% = t_{0.025}(9) \times \sqrt{MSE(1 + 1/b + 1/c + 1/bc)}; b = \text{No. of blocks}; C = \text{no. of checks}$$

$$= 2.262 \times \sqrt{1.23(1 + 1/4 + 1/4 + 1/16)} = 3.14 \text{ days}$$

8. FACTORIAL EXPERIMENTS

8.1 Simple Effects, Main Effects and Interaction

Factorial experiments are experiments in which two or more factors are studied together. Factor is a kind of treatment and in a factorial experiment any factor will supply several treatments. In factorial experiment, the treatments consist of combinations of two or more factors each at two or more levels.

Factorial experiment can be done in CRD, RCBD and Latin Square Design as long as the treatments allow. Thus, the term factorial describes specific way in which the treatments are formed and it does not refer to the experimental design used, *e.g.* nitrogen & phosphorus rates:

N = 0, 50, 100, 150 kg/ha

P = 0, 50, 100, 150 kg/ha

Kinds (noug cake, groundnut cake) and levels of protein supplement (25%, 50%, 75%).

The term level refers to the several treatments within any factor, *e.g.* if 5-varieties of sorghum are tested using 3-different row spacing the experiment is called 5×3 factorial experiment with 5 levels of variety factor (A) and three levels of spacing factor (B). An experiment involving 3 factors (variety, N-rate, weeding method) each at 2 levels is referred as $2 \times 2 \times 2$ or 2^3 factors; 3 refers to the number of factors and 2 refers to levels. Here we have 8 treatment combinations variety (x, y), N-rate (0, 50 kg/ha), weeding (with or without weeding). The $2^3 \times 3$ is a four factor experiment in which three factors each at 2-levels and the 4th factor at 3 levels.

If the above 2^3 factorial experiment is done in RCBD, the correct description of the experiment will be 2^3 factorial experiment in RCBD.

Interaction

Sometimes the factors act independent of each other. By this we mean that changing the level of one factor produces the same effect at all levels of another factor. Often, however, the effects of two or more factors are not independent. Interaction occurs when the effect of one factor changes as the level of the other factor changes, *e.g.* if the effect of 50kg N on variety x is 10 Q/ha and its effect on a variety y is 15 Q/ha, then there is interaction. When factors interact, the factors are not independent and a single factor experiment will lead to disconnected or misleading information. However, if there is no interaction it is concluded that the factors under consideration act independently of each other. Thus, results from separate single factor experiments are equivalent to those from a factorial experiment.

Example: A tall maize variety might out yield a short variety in high fertilizer rates due to high dry matter production.

Interaction is the failure of the differences in response to changes in levels of one factor to be the same at all levels of another factor or when the effect of one factor changes as the level of the other factor changes.

2 x 2 Factorial data of wheat yield (t/ha)

Variety (Factor A)	N-rate (kg/ha) (Factor B)		Simple effect of nitrogen on variety
	0 (b_0)	50 (b_1)	
X (a_0)	1.0	1.0	$(a_0b_1 - a_0b_0) = 0$
Y (a_1)	2.0	4.0	$(a_1b_1 - a_1b_0) = 2$
Simple effect of Variety	$(a_1b_0 - a_0b_0) = 1$	$(a_1b_1 - a_0b_1) = 3$	

Simple effects

- Simple effect of variety at N0: $2 - 1 = 1$
- Simple effect of variety at N1: $4 - 1 = 3$
- Simple effect of N on variety X: $1 - 1 = 0$
- Simple effect of N on variety Y: $4 - 2 = 2$

Main effects are the averages of the simple effects

- Main effect of variety = $\frac{1}{2}$ (Simple effect of A at b_0 + Simple effect of A at b_1)
 $= \frac{1}{2} [(a_1b_0 - a_0b_0) + (a_1b_1 - a_0b_1)] = \frac{1}{2} [(2 - 1) + (4 - 1)] = 2$
- Main effect of nitrogen = $\frac{1}{2}$ (Simple effect of factor B at a_0 + Simple effect of factor A at b_1) = $\frac{1}{2} [(a_0b_1 - a_0b_0) + (a_1b_1 - a_1b_0)] = \frac{1}{2} [(1 - 1) + (4 - 2)] = 1$

Interaction

It is calculated as the average of difference between simple effects of A at the two levels of B or the difference between the simple effects of B at the two levels of A.

$$\begin{aligned}
 &= \frac{1}{2} (\text{Simple effect of A at } b_1 - \text{simple effect of A at } b_0) \\
 &= \frac{1}{2} [(a_1b_1 - a_0b_1) - (a_1b_0 - a_0b_0)] = \frac{1}{2} [(4 - 1) - (2 - 1)] = 1 \\
 &\text{or} \\
 &= \frac{1}{2} (\text{Simple effect of B at } a_1 - \text{simple effect of B at } a_0) \\
 &= \frac{1}{2} (a_1b_1 - a_1b_0) - (a_0b_1 - a_0b_0) = \frac{1}{2} [(4 - 2) - (1 - 1)] = 1
 \end{aligned}$$

In factorial experiments, the following points should be noted:

- An interaction effect between two factors can be measured only if the two factors are tested together in the same experiment.
- When interaction is absent, the simple effect of a factor is the same for all levels of the other factors and equals to the main effect.
- When interaction is present, the simple effect of a factor changes as the level of the other factor changes.

Disadvantages of factorial experiments

- As the number of factors increase, the size of experiment becomes very large, *e.g.* with 8 factors each at 2-levels, there are 2^8 , 256 treatment combinations. Thus, experiments with this many treatments are costly to run.
- Large factorial experiments are difficult to interpret especially when there are interactions.

Uses of factorial experiments

- In exploratory experiments, where the aim is to examine a large number of factors to determine as which ones are important and which are not.
- To study relationships among several factors, to determine the presence and magnitude of interaction
- In experiments designed to lead to recommend over a wide range of conditions.

8.2 Two Factor Factorial in Randomized Complete Block Design

Example: An agronomist wanted to study the effect of different rates of phosphorus fertilizer on two varieties of common bean (*Phaseolus vulgaris*). He thought that the varieties might respond differently to fertilizer so he decided to use a factorial experiment with 2- factors: variety at two levels (T_1 = Indeterminate T_2 = Determinate) and phosphorus rate at 3 levels (P_1 = none, P_2 = 25 kg/ha, P_3 = 50 kg/ha).

Using the full factorial set of combinations, he had six treatments:

T_1P_1 ; T_1P_2 ; T_1P_3 ; T_2P_1 ; T_2P_2 ; T_2P_3 ,

He conducted this experiment using Randomized Complete Block Design with four blocks of six plots each.

Field layout and yield of common bean (Q/ha)

Block-I

T_2P_2	T_2P_1	T_1P_1	T_2P_3	T_1P_3	T_1P_2
8.3	11.0	11.5	15.7	18.2	17.1

Block-II

T_2P_1	T_2P_2	T_2P_3	T_1P_2	T_1P_1	T_1P_3
11.2	10.5	16.7	17.6	13.6	17.6

Block-III

T_1P_2	T_1P_1	T_2P_1	T_1P_3	T_2P_3	T_2P_2
17.6	14.3	12.1	18.2	16.6	9.1

Block-IV

T_1P_3	T_2P_2	T_2P_3	T_2P_1	T_1P_2	T_1P_1
18.9	12.8	17.5	12.6	18.1	14.5

The linear model for Two Factor Randomized Block Design:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \varepsilon_{ijk}$$

where, Y_{ijk} = the value of the response variable; μ = Common mean effect; α_i = Effect of factor A; β_j = Effect of block; γ_k = Effect of factor B; $\alpha\gamma_{ik}$ = Interaction effect of factor A & factor B; and ε_{ijk} = Experiment error (residual) effect

Steps of Analysis of Variance

1. Construct two way table for factors and calculate factor A total, Factor B total and grand total

Variety (Factor A)	Phosphorus (Factor B)			Factor A total (A)
	P1	P2	P3	
T1 (indeterminate)	53.9	70.4	72.9	197.2
T2 (determinate)	46.9	40.7	66.5	154.1
Factor B total (B)	100.8	111.1	139.4	351.3(G)

Block total				
Block	I	II	III	IV
Total	81.8	87.2	87.9	94.4

2. Using r as number of blocks, a as level of factor A, b level of factor B, compute C.F., total SS, block SS, treatment SS and Error SS

$$- \text{C.F.} = \frac{G^2}{rab} = \frac{(351.3)^2}{4 \times 2 \times 3} = 5142.15; \text{ where } r \text{ is number of replications, } a \text{ is level of factor A and } b \text{ is level of factor B}$$

$$- \text{Total SS} = (8.3)^2 + (11)^2 + \dots + (14.5)^2 - 5142.15 = 243.38$$

$$- \text{Block SS} = \frac{(81.8)^2 + (87.2)^2 + \dots + (94.4)^2}{6(ab)} - 5142.15 = 13.32$$

$$- \text{Treatment SS} = \frac{(53.9)^2 + (46.9)^2 + \dots + (66.5)^2}{4(r)} - 5142.15 = 221.38$$

$$- \text{Error SS} = \text{Total SS} - \text{Block SS} - \text{Treatment SS} = 243.38 - 13.32 - 221.38 = 8.68$$

3. Compute the three factorial components of treatment SS [partition treatments SS in to factor A SS, factor B SS, and A x B (interaction) SS]

$$- \text{Factor A (variety) SS} = \frac{\sum A^2}{rb} - \text{C.F.} = \frac{(197.2)^2 + (154.1)^2}{4 \times 3} - 5142.15 = 77.40$$

$$- \text{Factor B (P-rate) SS} = \frac{\sum B^2}{ra} - \text{C.F.} = \frac{(100.8)^2 + (111.1)^2 + (139.4)^2}{4 \times 2} - 5142.15 = 99.87$$

$$- \text{A} \times \text{B SS} = \text{Treatment SS} - \text{Factor A SS} - \text{Factor B SS} = 221.38 - 77.40 - 99.87 = 44.11$$

ANOVA TABLE

Source	DF	SS	MS	F-calcul.	F-table
--------	----	----	----	-----------	---------

					5%	1%
Block	r-1(4-1) = 3	13.32	4.44	7.65**	3.29	5.42
Variety (V)	a-1 (2-1) = 1	77.40	77.40	133.45**	4.54	8.68
Phosphorus (P)	b-1 (3-1) = 2	99.87	49.93	86.09	3.68	6.36
V × P	(a-1) (b-1) = 2	44.11	22.05	38.03	3.68	6.36
Error	(r-1) (ab-1) = 15	8.68	0.58			
Total	rab -1= 23	243.38				

$$CV = \frac{\sqrt{\text{Error } MS}}{\text{Grand mean}} \times 100; CV = \frac{\sqrt{0.58}}{14.64} \times 100 = 5.2\%$$

Interpretation of a factorial experiment

The interpretation of the results of factorial experiment depends on the outcome of the significance tests. If factor A × factor B interaction is significant, the main effects have no real meaning whether significant or not. In our case, since A × B interaction is highly significant, the results of experiment are best summarized in a two way table means of various A × B combinations. If interaction is not significant, then all of the information in the trial is contained in the significant main effects. In this case the results may be summarized in tables of mean for factors with significant main effects.

Mean Comparisons

There are three types of means in a two factor factorial experiment.

- Factor A means
- Factor B means
- Factor combinations (AB) or treatment means

Variety × Phosphorus rate means

	Phosphorus (B)			
Variety (A)	P1	P2	P3	Variety mean
T1 (indeterminate)	13.47	17.60	18.22	16.43
T2 (determinate)	11.72	10.17	16.62	12.84
Phosphorus mean	12.59	13.88	17.42	

Standard error of mean differences (\bar{Sd})

- \bar{Sd} to compare any two factor A means: $\bar{Sd} A = \sqrt{\frac{2 \times MSE}{rb}} = \sqrt{\frac{2 \times 0.58}{4 \times 3}} = 0.31 Q$
- \bar{Sd} to compare any two factor B means: $\bar{Sd} B = \sqrt{\frac{2 \times MSE}{ra}} = \sqrt{\frac{2 \times 0.58}{4 \times 2}} = 0.38 Q$
- \bar{Sd} to compare any two factor combination (treatment) means: $\bar{Sd} AB = \sqrt{\frac{2 \times MSE}{r}} = \sqrt{\frac{2 \times 0.58}{4}} = 0.54 Q$

8.3. Split-Plot Design

8.3.1 Uses, advantages and disadvantages

Split-plot design is frequently used for factorial experiments where the nature of experimental material makes it difficult to handle all factor combination. The principle underlying is that the levels of one factor are assigned at random to large experimental units. The large units are then divided into smaller units and then the levels of the second factor are assigned at random to small units within large units.

The large units are called the whole units or main-plots whereas the small units are called the split-plots or sub-plots (units). Thus, each main plot becomes a block for the sub-plot treatments. In split-plot design, the main plot factor effects are estimated from larger units, while the sub-plot factor effects and the interactions of the main-plot and sub-plot factors are estimated from small units.

As there are two sizes of experimental units, there are two types of experimental error, one for the main plot factor and the other for the sub-plot factor. Generally, the error associated with the sub-plots is smaller than that for the whole plots due to the fact that error degrees of freedom for the main plot are usually less than those for the sub-plots.

In split-plot design, the precision for the measurement of the effect of main plot factor is sacrificed to improve the precision of the measurement of the sub-plot factors.

Situations when to use split-plot design?

- a. When the level of one or more of the factors require larger amounts of experimental units than another. For instance, in field experiments, one of the factors could be method of land preparation (tractor, oxen, hand) and method of fertilizer application (broad cast, drill). These factors usually require larger experimental plots (units). The other factor could be varieties which can be compared using smaller units (plots). In this case methods of land preparation and fertilizer application can be assigned to main-plots and the varieties to the sub-plots.
- b. When an additional factor is to be incorporated in an experiment to increase its scope. For example, if the major purpose of an experiment is to compare the effect of several vaccines as a protectant against infection from certain disease of animals, to increase the scope of the experiment, several breeds of animals can be included which are known to differ in their

resistance to disease. Here, the breeds of animals could be arranged in main units and the vaccines to the subunits.

- c. When greater precision is desired for comparison of certain factors than others.

Since in a split-plot design, plot size and precision of measurement of the effects are not the same for both factors, the assignment of a particular factor to either the main-plot or to the sub-plot is extremely important.

Guidelines to apply factors either to main-plots or sub-plots:

- a. Degree of precision required: Factors which require greater degree of precision should be assigned to the sub-plot. For example, animal breeder testing three breeds of dairy cows under different types of feed stuff, will assign the breeds of animals to sub-units and the feed stuffs to the main unit. On the other hand, animal nutritionist may assign the feeds to the sub-units and the breeds of animals to the main-units as he is more interested on feed stuffs than breeds.
- b. Relative size of the main effect: If the main effect of one factor (factor A) is expected to be much larger and easier to detect than factor B, then factor A can be assigned to the main unit and factor B to the sub-unit. For instance, in fertilizer and variety experiments, the researcher may assign variety to the sub-unit and fertilizer rate to the main-unit, because he expects fertilizer effect to be much large and easier to detect than the varietal effect.
- c. Management practice: The factors, which require smaller amounts of experimental material, should be assigned to sub-plots. For example, in an experiment to evaluate the frequency of irrigation (5, 10, 15 days), on performance of different tree seedlings on nursery, the irrigation frequency factor could be assigned to the main plot and the different tree species to the sub-plots to minimize water movement to adjacent plots.

Advantages

- a. It permits the efficient use of some factors, which require large experimental units in combination with other factors, which require small experimental units.
- b. It provides increased precision in comparison of some of the factors (sub-plot factors).
- c. It promotes the introduction of new treatments into an experiment, which is already in progress.

Disadvantages:

- a. Statistical analysis is complicated because different factors have different error mean squares.
- b. Low precision for the main plot factor can result in large differences being non-significant, while small differences on the sub-plot factor may be statically significant even though they are of no practical significance.

8.3.2 Randomization and layout

There are two separate randomization process in split-plot design, one for the main plot factor and another for the sub-plot factor.

In each block, the main plot factors are first randomly applied to the main plots followed by random assignment of the sub-plot factors. Each of the randomization is done by any of the randomization schemes.

Example: An experiment was designed to test the effect of feeding four forage crops (Rhodes grass, Vetch, Alfalfa and Oat) on weight gain (kg/month) of the two breeds of cows (Zebu, Holstein). At the start of the experiment, it was assumed that breeds of cows would respond differently to the feed stuffs. Therefore, it was decided to use factorial experiment. The objective of the experiment was to compare the effect of forage crops as precisely as possible. Therefore, the experimenter assigned the breeds of animals to the main-plot and the four forage crops to the sub-plots. The experiment was replicated in three blocks (barns) based on initial body weight of animals as a blocking factor.

Procedures of randomization

Step 1: Divide the experimental area into $r = 3$ blocks, and divide each block into two main plots. Then randomly assign the two breeds of animals (H, Z) in each of the blocks.

Note that the arrangement of the main-plot factor can follow any of the designs: CRD, RCBD and LATIN square.

Step 2: Divide each of the main plot (unit) into 4-sub plots (units) and randomly assign the four feed stuffs (A, V, O, R) to each of the six-main plots (units).

Note:

Each main-plot factor is tested r -times where r is the number of blocks while each sub-plot factor is tested $a \times r$ times where a is level of factor A and r is the number of blocks. This is the primary reason for more precision for the sub-plot factors as compared to the main-plot factors.

The layout and the weight gain (kg/month) of the animals for feeding are given below:

Block I		Block II		Block III	
H	Z	H	Z	Z	H
A 25.9	R 15.5	O 18.0	V 22.7	O 13.2	V 28.4
V 25.3	A 18.9	A 26.7	O 13.5	A 19.6	A 27.6
O 19.3	O 13.8	V 24.8	R 15.0	V 22.3	R 25.4
R 22.2	V 21.0	R 24.2	A 18.3	R 15.2	O 20.5

- Main-plot factor is breed of animals: Holstein (H), Zebu (Z).
- Split-plot (units) factor is feed stuffs: Alfalfa (A), Vetch (V), Rhodes grass (R), Oat (O)

8.3.3 Analysis of variance

The linear model for Split Plot Design:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijk}$$

where, Y_{ijk} = the value of the response variable; μ = Common mean effect; α_i = Effect of factor A (main plot factor); β_j = Effect of block; γ_k = Effect of factor B (sub-plot factor); $(\alpha\beta)_{ij}$ = Interaction effect of factor A & Block (error a); $(\alpha\gamma)_{ik}$ = Interaction effect of factor A & B and ε_{ijk} = Experiment error (residual) effect (error b)

Steps of Analysis

Step 1: Arrange data by treatments (main-plot, sub-plot) and blocks and calculate main-plot total, and sub-plot total.

Treatments		Blocks		
Breeds	Feeds	I	II	III
Holstein	Alfalfa	25.9	26.7	27.6
	Vetch	25.3	24.8	28.4
	Oat	19.3	18.0	20.5
	Rhodes grass	22.2	24.2	25.4
Main-plot totals		92.7	93.7	101.9
Zebu	Alfalfa	18.9	18.3	19.6
	Vetch	21.0	22.7	22.3
	Oat	13.8	13.5	13.2
	Rhodes grass	15.5	15.0	15.2
Main-plot totals		69.2	69.5	70.3

Step 2: Construct two way tables of totals.

2.1. Block by factor A two-way table and compute block total, factor A total and grand total.

Weight gain totals of block by factor A.

Breeds (A)	Block 1	Block 2	Block 3	Factor A (breeds) total
Holstein	92.7	93.7	101.9	288.3
Zebu	69.2	69.5	70.3	209.0
Block total	161.9	163.2	172.2	
Grand total				497.3

2.2. Factor A by factor B total two-way table and calculate factor B totals

Breeds (A)	Feeds (B)			
	Alfalfa (b ₁)	Vetch (b ₂)	Oat (b ₃)	R. Grass (b ₄)
Holstein (a ₁)	80.2	78.5	57.8	71.8
Zebu (a ₂)	56.8	66.0	40.5	45.7
Factor B total	137.0	144.5	98.3	117.5

Step 3: Compute the correction factor and sum of squares for the main-plot analysis.

$$C.F. = \frac{G^2}{rab} = \frac{(497.3)^2}{3 \times 2 \times 4} = 10304.47$$

$$\text{Total SS} = (25.9)^2 + (25.3)^2 + \dots + (15.2)^2 - C.F. = 10820.59 - 10304.47 = 516.12$$

$$\begin{aligned} \text{Block SS} &= \frac{\sum (Block\ total)^2}{ab} - C.F. = \frac{(161.9)^2 + (163.2)^2 + (172.2)^2}{2 \times 4} - 10304.47 \\ &= 10312.34 - 10304.47 = 7.87 \end{aligned}$$

$$\begin{aligned} \text{Factor A (breeds) (main plot factor) SS} &= \frac{\sum (A\ total)^2}{rb} - C.F. = \frac{(288.3)^2 + (209)^2}{3 \times 4} - 10304.47 \\ &= 10566.49 - 10304.47 = 262.02 \end{aligned}$$

$$\text{Error (a) SS} = \frac{\sum (Block\ by\ A\ total)^2}{b} - C.F. - \text{Block SS} - \text{factor A SS}$$

$$\text{Error (a) SS} = \frac{(92.7)^2 + (69.2)^2 + \dots + (70.3)^2}{4} - 10304.47 - 7.87 - 262.02 = 5.03$$

Step 4: Compute the sum of squares for the sub-plot analysis.

$$\begin{aligned} \text{Factor B (feed stuff) SS} &= \frac{\sum (B\ total)^2}{ra} - C.F. \\ &= \frac{(137.0)^2 + (144.5)^2 + (98.3)^2 + (117.5)^2}{3 \times 2} - 10304.47 = 215.26 \end{aligned}$$

$$\begin{aligned} \text{Factor A} \times \text{Factor B SS} &= \frac{\sum (AB\ total)^2}{r} - C.F. - \text{Factor A SS} - \text{Factor B SS} \\ &= \frac{(80.2)^2 + (56.8)^2 + \dots + (45.7)^2}{3} - 10304.47 - 262.02 - 215.26 \\ &= 10800.45 - 10304.47 - 262.02 - 215.26 = 18.7 \end{aligned}$$

$$\begin{aligned} \text{Error (b) SS} &= \text{Total SS} - \text{total of all other sum of square} = \text{Total SS} - \text{Block SS} - \text{Factor A SS} - \text{Error (a) SS} - \text{Factor B SS} - \text{A} \times \text{B SS} \\ &= 516.12 - 7.87 - 262.02 - 5.03 - 215.26 - 18.7 = 7.24 \end{aligned}$$

Step 5: For each source of variation compute the mean squares by dividing the SS by its corresponding degrees of freedom.

$$\text{Block MS} = \frac{\text{Block SS}}{r-1} = \frac{7.87}{2} = 3.93$$

$$\text{Factor A MS} = \frac{\text{Factor A SS}}{a-1} = \frac{262.02}{1} = 262.02$$

$$\text{Error (a) MS} = \frac{\text{Error (a) SS}}{(r-1)(a-1)} = \frac{5.03}{2 \times 1} = 2.51$$

$$\text{Factor B MS} = \frac{\text{Factor B SS}}{b-1} = \frac{215.26}{3} = 71.75$$

$$A \times B \text{ MS} = \frac{AB \text{ SS}}{(a-1)(b-1)} = \frac{18.7}{3} = 6.23$$

$$\text{Error (b) MS} = \frac{\text{Error (b) SS}}{a(r-1)(b-1)} = \frac{7.24}{2 \times (3-1) \times (4-1)} = \frac{7.24}{12} = 0.60$$

Step 6: Compute the F-value for each effect that needs to be tested.

$$F(\text{block}) = \frac{\text{Block MS}}{\text{Error (a) MS}} = \frac{3.93}{2.51} = 1.56$$

$$F(A) = \frac{\text{Factor A MS}}{\text{Error (a) MS}} = \frac{262.02}{2.51} = 104.39$$

$$F(B) = \frac{\text{Factor B MS}}{\text{Error (b) MS}} = \frac{71.75}{0.60} = 119.58$$

$$F(A \times B) = \frac{A \times B \text{ MS}}{\text{Error (b) MS}} = \frac{6.23}{0.60} = 10.38$$

Step 7: Obtain the corresponding tabulated F-value and compare it with the calculated F-value at prescribed level of significance.

Source of variation	D.F.	SS	MS	Calculated F	Tabulated F	
					5%	1%
Block	(r-1) = 2	7.87	3.93	1.56	19.0	99.0
Breeds (A)	(a-1) = 1	262.02	262.02	104.39**	18.51	98.5
Error (a)	(r-1) (a-1) = 2	5.03	2.51			
Feeds (B)	(b-1) = 3	215.26	71.75	119.58**	3.49	5.95
A × B	(a-1) (b-1) = 3	18.70	6.23	10.38**	3.49	5.95
Error (b)	a(r-1)(b-1) = 12	7.24	0.60			
Total	(rab-1) = 23	516.12				

Step 8. Compute the two coefficients of variation, one corresponding to the main-plot analysis and another to the sub-plot analysis.

$$CV(a) = \frac{\sqrt{\text{Error (a) MS}}}{\text{Grand mean}} \times 100 = \frac{\sqrt{2.51}}{497.3/24} \times 100 = \frac{\sqrt{2.51}}{20.72} \times 100 = 7.6\%$$

$$CV(b) = \frac{\sqrt{\text{Error (b) MS}}}{\text{Grand mean}} \times 100 = \frac{\sqrt{0.60}}{20.72} \times 100 = 3.7\%$$

Note that CV (a) is greater than CV (b), this is because factors assigned to the main-plot are expected to be measured with less precision than that assigned to the sub-plot. If this trend does not frequently occur a competent statistician has to be consulted.

Mean comparisons:

Mean weight (kg/month) of Holstein and Zebu breeds fed with four types of forage species

Breeds (A)	Alfalfa (b ₁)	Vetch (b ₂)	Oat (b ₃)	R. Grass (b ₄)	Factor A mean
------------	---------------------------	-------------------------	-----------------------	----------------------------	---------------

					(A total/rb)
Holstein (a ₁)	26.7	26.2	19.3	23.9	24.0 (A ₁)
Zebu (a ₂)	18.9	22.0	13.5	15.2	17.4 (A ₂)
Factor B mean (B total/ra)	22.8 (B ₁)	24.1 (B ₂)	16.4 (B ₃)	19.6 (B ₄)	

Standard Errors of the Mean Differences

a. \bar{Sd} to compare two main plot factor (A) means (A₁ & A₂): $\bar{Sd} A = \sqrt{\frac{2 \times MSE(a)}{r \times b}}$ $\bar{Sd} A = \sqrt{\frac{2 \times 2.51}{3 \times 4}} = 0.65 \text{ kg}$

$LSD_{1\%} = t_{0.005} [\text{error (a) d. f.}] \times \bar{Sd} = LSD_{1\%} = t_{0.005} (2) \times 0.65 \text{ kg} = 9.925 \times 0.65 \text{ kg} = 6.45 \text{ kg}$

Since the mean difference (d) between A₁ and A₂ (24.0-17.4 = 6.6) > lsd value at 1% (6.45), the difference between the two means is highly significant.

b. \bar{Sd} to compare two sub-plot factor (B) means: $\bar{Sd} B = \sqrt{\frac{2 \times MSE(b)}{ra}} = \sqrt{\frac{2 \times 0.60}{3 \times 2}} = 0.45 \text{ kg}$

Example: To compare means of B₃ and B₂, $LSD_{1\%} = t_{0.005} [\text{error (b) d. f.}] \times \bar{Sd} = LSD_{1\%} = t_{0.005} (12) \times 0.45 \text{ kg} = 3.055 \times 0.45 = 1.37 \text{ kg}$.

Since the mean difference (d) between B₂ and B₃ (24.1 – 16.4 = 7.7) is greater than LSD value at 1% (1.37), the difference between the two means is highly significant.

c. \bar{Sd} to compare two sub-plot treatment means at the same level of main plot:

$$\bar{Sd} = \sqrt{\frac{2 \times MSE(b)}{r}} = \sqrt{\frac{2 \times 0.60}{3}} = 0.63 \text{ kg}$$

To compare (a₁b₁ = 26.7) with (a₁b₃ = 19.3); $LSD_{1\%} = t_{0.005} [\text{error (b) d. f.}] \times \bar{Sd} = LSD_{1\%} = t_{0.005} (12) \times 0.63 \text{ kg} = 3.055 \times 0.63 = 1.92 \text{ kg}$.

Since the mean difference (d) between (a₁b₁ & a₁b₃ = 26.7-19.3 = 7.4) is greater than LSD value at 1% (1.92 kg), the difference between the two means is highly significant.

d. \bar{Sd} to compare two sub-plot treatment means for at different main-plot factor level:

$$\bar{Sd} = \sqrt{\frac{2 \times [(b-1)MSE(b) + MSE(a)]}{rb}}$$

$$= \sqrt{\frac{2 \times [(4-1) \times 0.60 + 2.51]}{3 \times 4}} = 0.85 \text{ kg}$$

\bar{Sd} which contains both $MSE(a)$ and $MSE(b)$ has no exact value for the d.f. associated with it. To obtain an approximation:

$$\begin{aligned} \text{d.f.} &= \frac{[a(r-1)(a-1)][(b-1)MSE(b) + MSE(a)]^2}{[(a-1)(b-1)MSE(b)^2] + [a(MSE(a))^2]} \\ &= \frac{[2(2)(1)][(3 \times 0.60) + 2.51]^2}{[(1) \times 3 \times (0.60)^2] + 2(2.51)^2} = \frac{74.3}{1.08 + 12.60} = \frac{74.3}{13.68} = 5.4 \approx 5 \end{aligned}$$

To compare ($a_1b_1 = 26.7$) with ($a_2b_2 = 22.0$); $LSD_{1\%} = t_{0.005} [\text{error d.f.}] \times \bar{Sd} = LSD_{1\%} = t_{0.005}(5) \times 0.85 \text{ kg} = 4.032 \times 0.85 \text{ kg} = 3.43 \text{ kg}$.

Since the mean difference (d) between (a_1b_1 & $a_1b_3 = 26.7 - 22.0 = 4.7$) is greater than LSD value at 1% (3.43 kg), the difference between the two means is highly significant.

Presentation

If interaction of the factors is significant, results are summarized in two way table of means. However, if interaction is non-significant, the results are summarized in one way table of means.

9. COMPARISON OF TREATMENT MEANS

The F-test (ANOVA) shows whether there is significant difference among treatments or not. But, it does not show as which means are different from each other. There are many ways to compare the means of treatments tested in an experiment. One of these is pair comparison, the simplest and most commonly used comparisons in agricultural research.

There are two types of pair comparisons:

A. Planned pair comparison: In which the specific pair of treatments to be compared are identified before the start of the experiment, *e.g.* comparing the control treatment with each of the other treatments.

B. Unplanned pair comparison: In which no specific comparison is chosen in advance. Instead, every possible pair of treatment means are compared to identify pairs of treatments that are significantly different, *e.g.* variety trials.

The most commonly used test procedures for pair comparison in agricultural research are the Least Significant Difference and Tukey's test which are suitable for planned pair comparison and Duncan's Multiple Range Test (DMRT) which is applicable to an unplanned pair comparison.

9.1 Least Significant Difference (LSD) Test

The simplest and the most commonly used procedure for making pair comparisons. The procedure provides a single value at a prescribed level of significance, which serves as the boundary between significant and non-significant differences between any pair of treatment means. That is, two treatments are declared significantly different at a prescribed level of significance if their mean difference exceed the computed lsd value, otherwise they are not significantly different.

The LSD test is not valid for comparing all possible pair of means especially when the number of treatments is large. This is so because the number of possible pairs of treatment means increase rapidly as the number of treatments increase. In experiments where no real difference exists among all treatments, the numerical difference between the largest and smallest treatment means is expected to exceed the LSD value when the number of treatments is large.

To avoid this problem, the LSD test is used only when the F-test for treatment effect is significant and the number of treatments is not too large (less than six).

The procedure for applying the LSD test to compare any two treatments means

1. Rank the treatment means from the largest to the smallest in the column and from the smallest to largest in rows.
2. Compute all possible differences between the two treatment means to be compared.
3. Compute the LSD value at α level of significance

$$LSD_{\alpha} = t_{\alpha/2}(n) \times s\bar{d}$$

where $s\bar{d}$ = standard error of the treatment mean difference; $t_{\alpha/2}(n)$ is the table t-value at $\alpha/2$ level of significance and with n error degree of freedom

Example: Oil content (g) of linseed treated at different stages of growth with N-fertilizes tested in RCBD in four replications with error mean square of 1.31.

$$LSD_{5\%} = t_{0.025(15)} \times \sqrt{\frac{2MSE}{r}}, \text{ where MSE is error mean square; } r \text{ is the number of replications} = 2.131$$

$$\times \sqrt{\frac{2 \times 1.31}{4}} = 1.72 \text{ g}$$

$$LSD_{1\%} = t_{0.005(15)} \times \sqrt{\frac{2MSE}{r}} = 2.947 \times \sqrt{\frac{2 \times 1.31}{4}} = 2.39 \text{ g}$$

4. Compare the mean difference (d) in step 2 with LSD value computed in step (3) using the following rule:

- if $|d| > LSD$ value at 1% level of significance, there is highly significant difference between the two treatment means compared (put two asterisks on differences).
- if $|d| > LSD$ value at 5% level of significance but $\leq LSD$ value at 1% level of significance, there is significant difference between the two treatment means compared (put one asterisks on differences)
- if $|d| \leq LSD$ value at 5% level of significance, the two treatment means compared are not significantly different (put n.s.)

No	Treatments (stage of application of N)	Treatment mean (g)
1	Seedling	5.10
2	Early blooming	4.30
3	Half blooming	4.00

4	Full Blooming	6.70
5	Ripening	6.05
6	Unfertilized (control)	7.03

Treatments	4.00 (T3)	4.30 (T2)	5.10 (T1)	6.05 (T5)	6.70 (T4)	7.03 (T6)
7.03 (T6)	3.03**	2.73**	1.93*	0.98 ^{ns}	0.33 ^{ns}	-
6.70 (T4)	2.70**	2.40**	1.60 ^{ns}	0.65 ^{ns}	-	
6.05 (T5)	2.05*	1.75*	0.95 ^{ns}	-		
5.10 (T1)	1.10 ^{ns}	0.80 ^{ns}	-			
4.30 (T2)	0.30 ^{ns}	-				
4.00 (T3)	-					

Thus, the differences between T6 & T3, T6 & T2, T4 & T3, T4 & T2 are highly significant; while the differences between T3 & T5, T1 & T6, T2 & T5 are significant.

Note that there are $\frac{t(t-1)}{2}$ possible (unplanned) pair comparisons and (t-1) planned pair comparisons where t is the number of treatments. In the above example, 15 unplanned pair comparisons and five planned pair comparisons are possible.

Presentation of data using LSD

Table __. Mean oil content of linseed treated with nitrogen fertilizer at different stages

Stage of application	Oil content (g)
Seedling	5.10
Early blooming	4.30
Half-blooming	4.00
Full- blooming	6.70
Ripening	6.05
Unfertilized	7.03
LSD _(0.05)	1.72 g
CV (%)	20.7

9.2 Duncan's Multiple Range Test (DMRT)

It is most widely used to make all possible pair comparisons. The procedure for applying the DMRT is similar to LSD test but it requires progressively larger values for significance between the treatment means as they are more widely separated in the array.

The test is more appropriate when the total number of treatments is large. It involves the calculation of the shortest significant difference (SSD).

The SSD is calculated for all possible relative positions (P) between the treatment means when the means are arranged in order of magnitude (in decreasing or increasing order).

Procedure

Step 1: Arrange all the treatment means in increasing or decreasing order. Data such as milk & crop yield are usually arranged from the highest to the lowest.

Example: Yields (kg/plot) of wheat varieties grown in 4 by 4 Latin Square Design with error mean square of 0.45:

B (12.3) A (12.00) C (10.8) D (6.7)

Step 2: Calculate $s\bar{d}$ (the standard error of the treatment mean difference) as:

$$s\bar{d} = \sqrt{\frac{2MSE}{r}} = s\bar{d} = \sqrt{\frac{2 \times 0.45}{4}} = 0.47 \text{ kg}$$

Step 3: Calculate the shortest significant difference (SSD) for relative positions (P) in the array of means. Since we have four treatment means they can be 2, 3 and 4 distance apart. B and A are 2 distance apart (P = 2); B and C are 3 distance apart (P = 3); B and D are 4 distance apart (P = 4); A and D are 3 distance apart (P = 3); etc.

For the above example, the R values with error d. f. of 6 at 1% level of significance are found from R-table.

P =	2	3	4
R _{0.01} =	5.24	5.51	5.65
SSD =	1.74	1.83	1.88

P = the distance in ranks between the pairs of treatment means to be compared.

R = significant student size range at error d. f. (6).

$$\text{SSD (Shortest Significant Difference)} = \frac{R \times s\bar{d}}{\sqrt{2}}$$

$$\text{SSD at } P = 2 = \frac{5.24 \times 0.47}{\sqrt{2}} = 1.74$$

$$\text{SSD at } P = 3 = \frac{5.51 \times 0.47}{\sqrt{2}} = 1.83$$

$$\text{SSD at } P = 4 = \frac{5.65 \times 0.47}{\sqrt{2}} = 1.88$$

Note that SSD values increase as the distance between treatments (P) to be compared increases.

Step 4: Test the difference between treatment means in the following order.

- Largest – Smallest = $12.3 - 6.7 = 5.6$ compare with SSD value at $(P = 4) = 1.88$; $d(5.6) > SSD$ at $P = 4$ (1.88); thus the difference is significant at 1% level of significance.
- Largest – 2nd smallest = $12.3 - 10.8 = 1.5$ compare with SSD at $(P=3) = 1.84$; $d(1.5) < SSD$ at $P=3$ (1.84); thus, the difference is non-significant at 1% level of significance.
- Largest – 2nd largest = $12.3 - 12.0 = 0.3 < SSD$ at $P = 2$ (1.75); thus, the difference is non-significant at 1% level of significance.
- 2nd largest – smallest = $12.0 - 6.7 = 5.3$ compared with SSD at $P = 3$ (1.84); $d(5.3) > SSD$ (1.84) at $P = 3$; significant at 1% level of significance
- 2nd smallest – smallest = $10.8 - 6.7 = 4.1$ compared with SSD at $P = 2$ (1.75); $d(4.1) > SSD$ (1.75) at $P = 2$; significant
- etc

Note that SSD value at $(P=2)$ is equals to LSD value

	B (12.3)	A (12.0)	C (10.8)	D (6.7)
D (6.7)	5.6 ^{**} (P=4)	5.3 ^{**} (P=3)	P = 4.1 ^{**} (P=2)	-
C (10.8)	1.5 ^{ns} (P=3)	1.2 ^{ns} (P=2)	-	
A (12.0)	0.3 ^{ns} (P=2)	-		
B (12.3)	-			

SSD ($P = 4$) = 1.88; SSD ($P = 3$) = 1.83; SSD ($P = 2$) = 1.74

Treatments B & D, A & D, C & D are significantly different at 1%, while treatments B & C, B & A, and A & C are not significantly different at 1% level of significance.

Step 5: Present the test result in one of the following two ways.

- A. Use a line notation if the sequence of results can be arranged according to their ranks.
- B. Use the alphabet notation if the desired sequence of the results is not based on their rank which is commonly used.

Any two means underscored by the same line are not significantly different at 1% level of significance according to DMRT.

B(12.3) A(12.0) C(10.8) D(6.7)

The alphabet notation can be derived from line notation simply by assigning the same alphabet to all treatment means connected by the same horizontal line. It is usual practice to assign letter a for the first line, b for second line, c for third and so on. Note that letter a can be for the largest or smallest treatment mean depending on the rank of arrangement.

Presentation of data using DMRT

Table __ Mean yields of wheat varieties planted at Debrezeit Agricultural Research Center.

Variety	Yield (kg)
A	12.0 a
B	12.3 a
C	10.8 a
D	6.7 b

Note that we have to put a footnote below the table stating that any two means in the same column followed by the same letter are not significantly different at 1% level of significance according to DMRT. Note also that both LSD and DMRT are not used in the same table. Use either of them depending on the appropriateness of the test.

9.3 Tukey's Test

It is more conservative than LSD test because it requires the largest treatment mean differences for significance.

It is computed in a manner similar to the LSD test except standard error of the mean is used instead of standard error of the mean difference ($s\bar{d}$).

Studentized range (q-table) is used in place of t-table

The procedure involves:

1. Select a value from q table, which depends on number of means (n) and error degree of freedom (v).
2. Compute the Critical Difference (CD) as $= q_{\alpha(n, v)} \times \sqrt{\frac{MSE}{r}}$ where MSE is error mean square; n is number of means to be compared; v is error degrees of freedom and r is number of replications.
3. For any pair of means if the absolute value of the difference $|d| > \text{critical value}$, the difference is judged to be significant at a prescribed level of significance.

Example: The following analysis of variance table is from CRD with six varieties replicated four times in glass house (mean rust incidence)

Source	d. f.	MS	F-cal.	F-table (5%)
Variety	(t-1) = 5	2976.44	24.80**	2.77
Error	t(r-1) = 18	120.00		

Variety:	1	2	3	4	5	6
Mean stem rust incidence (%):	50.3	69.0	24.0	94.0	75.0	95.3

n= 6; V= 18; $q_{0.05}(6, 18) = 4.495$

$$CD = q_{\alpha} \times \sqrt{MSE / r} = 4.495 \times \sqrt{\frac{120}{4}} = 24.62\%$$

Difference between means

	24.0(3)	50.3(1)	69.0(2)	75(5)	94(4)	95.3(6)
95.3(6)	71.3*	45.0*	26.3*	20.3 ^{ns}	1.3 ^{ns}	-
94(4)	70.0*	43.7*	25.0*	19.0 ^{ns}	-	
75(5)	51.0*	24.7*	6.0 ^{ns}	-		
69(2)	45.0*	18.7 ^{ns}	-			
50.3(1)	26.3*	-				
24.0(3)	-					

Thus, differences between varieties 6&3, 4&3, 5&3, 2&3, etc. are significant while differences between varieties 2&1, 5&2, etc. are non-significant.

9.4. Pair Comparisons with Missing Data

In applying the LSD test and DMRT, it is important that the appropriate standard error of the mean difference ($s\bar{d}$) should be used. $s\bar{d}$ is affected by the experimental design used, the number of replications of the two treatments being compared, and the specific type of means to be compared.

A). In CRD, RCBD and Latin Square Design where the number of replications for all treatments is equal, the $s\bar{d}$ for any pair of treatment means is computed as:

$$s\bar{d} = \sqrt{\frac{2MSE}{r}}$$

Where, MSE is mean square for error; r = number of replications that is common to all treatments.

Thus, $lsd_{\alpha} = t_{\alpha/2}(n) \sqrt{\frac{2 \times MSE}{r}}$ where n is error degree freedom.

B). When the two treatments do not have the same number of replications in CRD. $s\bar{d}$ is computed as:

$$s\bar{d} = \sqrt{MSE \left(\frac{1}{r_i} + \frac{1}{r_j} \right)}$$

where MSE is mean square error; r_i and r_j are the number of replications of the two treatment means (i & j) to be compared.

$$\text{Thus, } LSD_{\alpha} = t_{\alpha/2}(n) \sqrt{MSE \left(\frac{1}{r_i} + \frac{1}{r_j} \right)}$$

Example: CRD with an unequal replications, effect of 4 – types of feedstuff on weight gain of chicks.

Treatment A = given to 5- chicks (5-replications) = 43.8 g

Treatment B = given to 4 chicks (4 replications) = 73.0 g

Treatment C = given to 3 chicks (3 replications) = 73.33 g

Treatment D = given to 5 chicks (5 replications) = 142.8 g

To compare treatment B with treatment D:

Given error mean square of 843.1 and error degree of freedom of 13, test if there is significant difference between treatments B & D.

Difference between treatment mean: $(D - B) = d = 142.8 - 73.0 = 69.8g$

$$LSD_{5\%} = t_{0.025}(13) \times \sqrt{MSE \left(\frac{1}{r_B} + \frac{1}{r_D} \right)} = 2.160 \times \sqrt{843.1 \times \left(\frac{1}{4} + \frac{1}{5} \right)} = 42.07g$$

$$LSD_{1\%} = t_{0.005}(13) \times \sqrt{MSE \left(\frac{1}{r_B} + \frac{1}{r_D} \right)} = 3.012 \times \sqrt{843.1 \times \left(\frac{1}{4} + \frac{1}{5} \right)} = 58.67g$$

Compare treatment mean difference $|d|$ with the calculated lsd value. Since $|d|$ 69.8 > LSD value at 1% (58.67), there is a highly significant difference between treatments B and D or treatment D significantly increased the weight of chicks as compared to treatment B.

C). $s\bar{d}$ for the treatments with a single missing value and that of any other treatment without missing values.

a) For RCBD: $s\bar{d} = \sqrt{MSE \left[\frac{2}{r} + \frac{t}{r(r-1)(t-1)} \right]}$. Thus, $LSD_{\alpha} = t_{\alpha/2}(\text{error d.f.}) \times s\bar{d}$ where MSE is mean square of error; t = no. of treatments; r = no. of replications.

- b) Latin Square Design: $s\bar{d} = \sqrt{MSE \left[\frac{2}{r} + \frac{1}{(t-1)(t-2)} \right]}$. Thus, $LSD_{\alpha} = t_{\alpha/2}(\text{error d.f.}) \times s\bar{d}$ where MSE is mean square error of the analysis of variance of Latin Square Design with a single missing value; r = number of replications.

10. ANALYSIS OF COVARIANCE

The analysis of covariance simultaneously examines the variance and covariance of selected variables so that the character of primary interest is more accurately characterized than by the use of analysis of variance only. Analysis of covariance requires measurement of the character of interest and the measurement of one or more variable(s) known as covariate(s). It also requires that the functional relationship of the co-variates (x) with the character of primary interest (y) is known before hand.

Examples: Consider wheat variety trial in which weed infestation is used as a co-variate with a known functional relationship between weed incidence and grain yield (the character of primary interest), the covariance analysis can adjust grain yield in each plot to a common level of weed incidence. With the covariance analysis, the variation in yield due to weed incidence is quantified and effectively separated from that due to varieties, *e.g.* age or initial body weight of experimental animals can be used as a covariate and weight gain due to rations as character of interest.

Covariance analysis can be applied to any number of covariates and to any type of functional relationships between variables. In this section, however, we will deal with the case of a single covariate whose relationships to character of primary interest is linear.

10.1 Uses of Covariance Analysis

1. To Control Experimental Error

One way to reduce experimental error is by using proper type of blocking. However, blocking can not cope with certain types of variability such as spotty soil heterogeneity and unpredictable insect/disease incidence. In such cases heterogeneity between experimental plots does not follow a definite pattern. Thus, use of covariance analysis should be considered in experiments in which blocking can not adequately reduce the experimental error. By measuring an additional variable (*i.e* covariate) that is known to be linearly related to the primary variable (y) the source of variation associated with the covariate can be deducted from experimental error.

The experimental error is reduced and the precision for comparing treatment increased, *e.g.* in a cattle feeding experiment to compare the effects of several rations on weight gain, animals assigned to any one block may vary in initial weight. Now if the initial weight is correlated with gain in weight, a portion of experimental error for gain can be the result of differences in initial weight. By covariance analysis, a contribution, which can be attributed to differences in initial weight, can be computed and eliminated from experimental error.

2. Adjustment of Treatment Mean

With the covariance analysis, the primary variable (y) can be adjusted linearly up wards or down wards depending on the relative size of its respective covariate (x). The treatment means of dependent variable is adjusted to a value that it would have had there been no differences in the value of co-variate, *e.g.* age of cows, initial body weight of the cow, etc. In situations where real differences among treatments for the independent variable do occur but are not the direct effect of the treatments, adjustment is warranted, *e.g.* in variety trial, if seeds may differ widely in germination, not because of inherent differences of the varieties, adjustment can be done, but if the density is treatment by itself there is no need to adjust the means.

3. Interpretation of Experimental Results

Covariance analysis aids the experimenter in understanding the principles underlying the results of an investigation. By examining the primary character of interest (y) together with other characters (x) whose functional relationship to y are known, the biological process governing the treatment effects on y can be characterized more clearly.

4. Estimation of Missing Data

The missing data formula technique biases the treatment sum of squares upwards. The use of covariance analysis to estimate the missing value(s) results in a minimum residual sum of squares and unbiased treatment sum of squares.

10.2. Computation Procedure

Covariance analysis is an extension of analysis of variance. It is a combination of analysis of variance and linear regression. Covariance analysis can be used for CRD; RCBD and split-plot designs, but the computation procedures vary some what.

Computation procedure for RCBD

The following data show ascorbic acid content of ten varieties of common bean. From the previous experience it was known that increase in maturity resulted in decrease in vitamin C content. Since all varieties were not of the same level of maturity on the same day, it was not possible to harvest all plots at the same stage of maturity. Hence, the percentage of dry matter based on 100 g of freshly harvested beans was observed as an index of maturity and used as a covariate.

Ascorbic acid content (ASAC, mg/100 g of seed) and percentage of dry matter (% DM) for common bean varieties:

Variety	Block I		Block 2		Block 3	
	% DM	ASAC	% DM	ASAC	% DM	ASAC
1	34	93	33	95	35	92
2	40	47	40	51	51	33
3	32	81	30	100	34	72
4	38	67	38	74	40	65

5	25	119	24	128	25	125
6	30	106	29	111	32	99
7	33	106	34	107	35	97
8	34	61	31	83	31	94
9	31	80	30	106	35	77
10	21	149	25	151	23	170

Conduct the analysis of covariance & calculate standard error of mean difference.

Treatment	% DM	ASA C	% DM	ASA C	% DM	ASA C	Treatment total	
	X	Y	X	Y	X	Y	X	Y
1	34	93	33	95	35	92	102	280
2	40	47	40	51	51	33	131	131
3	32	81	30	100	34	72	96	253
4	38	67	38	74	40	65	116	206
5	25	119	24	128	25	125	74	372
6	30	106	29	111	32	99	91	316
7	33	106	34	107	35	97	102	310
8	34	61	31	83	31	94	96	238
9	31	80	30	106	35	77	96	263
10	21	149	25	151	23	170	69	470
Block total	318	909	314	1006	341	924	973	2839

Steps of Analysis

1. Conduct analysis of variance for each of the variables, covariance and sum of treatment and error sum of squares

Source	D. F.	SS of ASAC (Y)	SS of % DM (X)	SS of XY
Block	2	545.3	42.47	-75.23
Treatment	9	25689.0	972.70	-4633.23
Error	18	1608.7	86.20	-251.77
Treatment + Error	27	27297.7	1058.90	-4885.00

2. Analyse covariance

$$C.F. = \frac{G_x \times G_y}{r \times t} = \frac{973 \times 2839}{3 \times 10} = 92078.23$$

$$\text{Total Sum of Products} = \sum x_i \times y_i - C.F. = (34 \times 93) + (40 \times 47) + \dots + (23 \times 170) - 92078.23 = 87118 - 92078.23 = -4960.23$$

$$\text{Sum of Products due to Blocks} = \frac{\sum B_x \times B_y}{t} - \text{C.F.} = \frac{(318 \times 909) + (314 \times 1006) + (341 \times 924)}{10} -$$

$$92078.23 = 92003 - 92078.23 = -75.23$$

$$\text{Sum of Products for Treatments: } \frac{\sum T_x \times T_y}{r} - \text{C.F.} = \frac{(102 \times 280) + (131 \times 131) + \dots + (69 \times 470)}{3} -$$

$$92078.23 = -4633.23$$

$$\text{Error Sum of Squares of Products} = \text{Total Sum of Products} - \text{Block SS of Products} - \text{Treatment SS of Products} = -4960.23 - (-75.23) - (-4633.23) = -251.77$$

$$3. \text{ Compute the adjusted error SS of Y as} = \text{Error SS due to Y} - \frac{(\text{Error SS of } xy)^2}{\text{Error SS of } x}$$

$$= 1608.7 - \frac{(-251.77)^2}{86.2} = 873.34$$

$$4. \text{ Compute (treatment + error) adjusted SS of Y as:}$$

$$= (\text{Treatment} + \text{Error SS of Y}) - \frac{(\text{Treatm.} + \text{Error SS of } xy)^2}{\text{Treatm.} + \text{Error SS of } x}$$

$$= 27297.7 - \frac{(-4885)^2}{(1058.9)} = 4761.84$$

$$5. \text{ Compute the treatments adjusted SS of Y} = \text{Treatment} + \text{Error adjusted SS of Y} - \text{Error adjusted SS of Y} = 4761.84 - 873.34 = 3888.5$$

Source	D.F	SS	MS	F-cal	F table (1%)
Treatment (adjust)	(t-1) 10-1 = 9	3888.50	432.0	8.41**	3.68
Error (adjusted)	(r-1) (t-1)-1 = 17	873.34	51.4		

$$6. \text{ Compute regression coefficient or slope of data}$$

$$= \frac{\text{Error Sum of Products}}{\text{Error SS due to } x} = \frac{-251.77}{86.2} = -2.92$$

$$7. \text{ Compute adjusted treatment mean as: } \bar{Y} \text{ adjusted} = \bar{Y}_i(\text{unadjusted}) - b \times (\bar{X}_i - \bar{X}_{\text{grand mean}}). \text{ For example, for treatment 1= } \bar{Y} \text{ adjusted} = 93.33 - (-2.92 \times 1.57) = 97.92$$

No	Average ASAC (Unadjusted)	Average % DM (2)	Deviation of % DM ($\bar{x}_{\text{mean}} - \bar{x}_{\text{grand mean}}$)	Adjusted ASAC
1	93.33	34.00	1.57	97.92
2	43.67	43.67	11.24	76.48
3	84.33	32.00	-0.43	83.08
4	68.67	38.67	6.24	86.88
5	124.00	24.67	-7.76	101.33
6	105.33	30.33	-2.10	99.21
7	103.33	34.00	1.57	107.92
8	79.33	32.00	-0.43	78.08

9	87.67	32.00	-0.43	86.41
10	156.67	23.00	-9.43	129.13

Compute the relative efficiency (R.E.) of covariance analysis compared to standard analysis of variance

$$R.E. = \frac{(Error\ unadjusted\ MS\ of\ Y)}{(Error\ adjusted\ MS\ of\ Y)(1 + \frac{Treat\ MS\ of\ X}{Error\ SS\ of\ X})} \times 100 = \frac{(1608.7/18)}{(51.4) \times (1 + \frac{972.7/9}{86.2})} \times 100 = \frac{89.37}{115.84} \times 100 = 77.15\%$$

Thus, the result indicates that the use of % dry matter as the covariate has not increased precision in ascorbic acid content which would have been obtained had the ANOVA is done without covariance.

$$CV = \frac{\sqrt{Adjusted\ Error\ MS\ of\ Y}}{Grand\ Mean\ of\ Y} \times 100 = \frac{\sqrt{51.4}}{94.63} \times 100 = 7.6\%$$

Mean comparison

\bar{Sd} to compare two adjusted treatment means:

$\sqrt{Adjusted\ Error\ Mean\ Square\ of\ Y[\frac{2}{r} + \frac{(x_i - x_j)^2}{Error\ SS\ of\ X}]}$ where x_i & x_j are the covariate means of i^{th} and j^{th} treatment; r is the number of replications common to both treatments. For instance, to compare means of T1 & T2:

$\sqrt{51.4[\frac{2}{3} + \frac{(43.67 - 34)^2}{86.2}]}$ where 34 & 43.67 are the covariate means of 1st and 2nd treatments; 3 is the number of replications common to both treatments.

$$\bar{Sd} = \sqrt{90.02} = 9.49$$

11. Combine Analysis of Data

In field experiments, it is necessary to repeat the experiments over a number of locations, seasons, or both, *e.g.* varietal trials, plant spacing, fertilizer trials. The purpose of repeating the experiments is to find recommendation that can be applied over space (location), time (season) or both.

In such repeated experiments, appropriate statistical procedures for a combined analysis of data have to be used. The main purposes of combined analysis of data are:

- to estimate the average response to a given experiment
- to test the consistency of the response from place to place or year to year, *i.e.* to determine if there is interaction effect of the treatments, *e.g.* stability analysis of varieties.

If the response is consistent from place to place and year to year, it shows the absence of interaction.

Steps of combined analysis

1. Construct an outline of combined analysis over years or locations or both on the basis of experimental design used. For example, the outline for the experiment conducted at six environments with five treatments and six replications in RCBD the outline of ANOVA is given as:

Source	Degrees of freedom	Mean squares	Computed F
Environment (E)	(e-1) = 5	EMS	EMS/RMS
Blocks/within environment	e(r-1) = 30	RMS	
Treatments (T)	(t-1) = 4	TMS	TMS/MSE
T x E	(e-1) (t-1) = 20	ExTMS	E xTMS/MSE
Pooled error	e(r-1) (t-1) = 120	MSE	

Where e = no. of environments; r = no. of replications; t = no. of treatments; EMS = environment mean square; RMS is blocks/replications mean square; TMS = treatment mean squares; and MSE = mean square error.

2. Compute the usual ANOVA for each environment according to RCBD.

Environment:	E1	E2	E3	E4	E5	E6
Error d.f. (r-1)(t-1)	20	20	20	20	20	20
Error MS	5776	4028	4516	9526	7056	5535

3. Test the homogeneity of the variances

Case 1: When there are only two environments, use F-test as:

F calculated = $\frac{\text{Larger error mean square}}{\text{Smaller error mean square}}$ and compare with table F-value at d.f. for larger error mean square as numerator and d.f. for smaller error mean square as denominator. If F-calculated is > F-table, the null hypothesis of homogeneity of variances is rejected, i.e. the error variances are heterogeneous. On the other hand, if the F-calculate value is \leq F-table, the error variances are homogeneous, we can analyze the data separately or transform the data to homogenize the error variances to use combined analysis.

Case 2: When environment is more than two, use Bartlett's chi-square test

$$\chi^2_c = \frac{2.3026 \left(\sum_{i=1}^k k_j \log \bar{MSE} - \sum_{i=1}^k k_j \log MSE_i \right)}{1 + \left(\frac{1}{3(e-1)} \right) \left(\sum_{j=1}^k \frac{1}{k_j} - \frac{1}{\sum k_j} \right)}$$

Where

K_j = Error d.f. for each environment

\bar{MSE} = Pooled error mean square = $\frac{\sum K_j \times MSE_i}{\sum K_j} = \frac{20 \times 5776 + 20 \times 4028 + \dots + 20 \times 5535}{120} = 6072.83$

MSE_i = Error mean square for each environment

e = number of environment

Environment	K_j	MSE	$\log MSE$	$K_j \times \log$	$K_j \times \log$	$1/k_j$
-------------	-------	-----	------------	-------------------	-------------------	---------

				pooled MSE	MSEi	
E1	20	5776	3.761627	75.6	75.23254	0.05
E2	20	4028	3.605089	75.6	72.10179	0.05
E3	20	4516	3.654754	75.6	73.09508	0.05
E4	20	9526	3.978911	75.6	79.57821	0.05
E5	20	7056	3.848559	75.6	76.97117	0.05
E6	20	5535	3.743118	75.6	74.86235	0.05
				453.6	451.8411	0.3

$$\bar{MSE} = \text{Pooled error mean square} = \frac{\sum K_j \times MSE_i}{\sum K_j} = \frac{20 \times 5776 + 20 \times 4028 + \dots + 20 \times 5535}{120} = 6072.83$$

$$\text{Log}(6072.83) = 3.78$$

$$\chi^2_c = \frac{2.3026(453.6 - 451.84)}{1 + \left(\frac{1}{3(5)}\right)(0.3 - 0.008)}; \chi^2_c = 3.97$$

$$\chi^2_{(c-1) 5\%} = \chi^2_{(5) 5\%} = 11.07$$

Since χ^2_c (3.97) is less than χ^2 table at 5% (11.07), accept the null hypothesis. Thus, this implies the errors are homogeneous

12. REGRESSION AND CORRELATION ANALYSIS

12.1 Types of Regression & Correlation

Regression analysis describes the effect of one or more variables (designated as independent variables) on a single variable (designated as the dependent variable). It expresses the dependent variable as a function of independent variable(s).

For regression analysis, it is important to clearly distinguish between the dependent and independent variables.

Examples:

- Weight gain in animals depends on feed
- Number of growth rings in a tree depends on age of the tree
- Grain yield of maize depends on a fertilizer rate

In the above cases, weight gain, number of growth rings and grain yield are dependent variables, while feed, age and fertilizer rates are independent variables.

The independent variable is designated by x and the dependent variable by y.

Correlation analysis, on the other hand, provides a measure of the degree of association between the variables, e.g. the association between height and weight of students; body weight of cows and milk production; grain yield of maize and thousand kernel weight.

Regression and correlation analysis can be classified:

- a. Based on the number of independent variables as:
 - Simple: one independent variable and one dependent variable.
 - Multiple: if more than one independent variables and a dependent variable is involved
- b. Based on the form of functional relationship as:
 - Linear: if the form of underlying relationship is linear
 - Non-linear: if the form of the relationship is non-linear

Thus, regression and correlation analysis can be classified into 4:

- Simple linear regression and correlation analysis
- Multiple linear regression and correlation analysis
- Simple non-linear regression and correlation analysis
- Multiple non-linear regression and correlation analysis

Linear Relationships

The relationship between any two variables (independent and dependent) is linear if the change in y is constant as x changes through out the range of x under consideration.

The functional form of linear relationship between a dependent variable y and an independent variable x is represented by the equation.

$$y = a + bx$$

y = is the dependent variable

a = the intercept of the line on the y-axis (the value of y when x is 0)

b = linear regression coefficient, is the slope of the line or the amount of change in y for each unit change in x.

When there are more than one independent variables as say k-independent variables (x_1, x_2, \dots, x_k), the simple linear regression equation $y = a + \beta x$ can be extended to the multiple linear functional form of:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

where α is the y intercept (the value of y when all x's are 0); $\beta_1, \beta_2, \dots, \beta_k$ are partial regression coefficients associated with the independent variables.

12.2 Simple Linear Regression and Correlation Analysis

12.2.1 Simple linear regression analysis

The simple linear regression analysis deals with the estimation and test of significance concerning the two parameter α and β in the equation:

$$Y = \alpha + \beta_x$$

The data required for the application of the simple linear regression are the n-pairs (with $n > 2$) of y and x values.

Steps to estimate α and β

Step 1: Compute the means (\bar{x} and \bar{y}), deviation from means [$(x - \bar{x})$, $(y - \bar{y})$], square of the deviates (x^2 , y^2) and product of deviates (xy).

Example: Determine the regression equation for dependence of wing length of 13 sparrows of various ages.

Age (days), x	Wing length (cm), Y	Deviations from the mean		Square of deviates		Product of deviates
		$x - \bar{x}(x)$	$y - \bar{y}(y)$	x^2	y^2	
3	1.4	-7	-2.015	49	4.060225	14.105
4	1.5	-6	-1.915	36	3.667225	11.49
5	2.2	-5	-1.215	25	1.476225	6.075
6	2.4	-4	-1.015	16	1.030225	4.06
8	3.1	-2	-0.315	4	0.099225	0.63
9	3.2	-1	-0.215	1	0.046225	0.215
10	3.2	0	-0.215	0	0.046225	0
11	3.9	1	0.485	1	0.235225	0.485
12	4.1	2	0.685	4	0.469225	1.37
14	4.7	4	1.285	16	1.651225	5.14
15	4.5	5	1.085	25	1.177225	5.425
16	5.2	6	1.785	36	3.186225	10.71
17	5	7	1.585	49	2.512225	11.095
$\Sigma = 130$	44.4			262	19.6569	70.8
Mean = 10.0	3.415					

Step 2: Compute the estimates of the regression parameter α and β as:

$\bar{y} = a + b\bar{x}$, where a is the estimate of α (the y intercept) and b is the estimate of β (linear regression coefficient, slope).

$$b = \frac{\sum xy}{\sum x^2} = \frac{70.8}{262} = 0.27 \text{ cm/day} ; a = \bar{y} - b\bar{x}$$

$$a = 3.415 - 0.27 \times 10 = 0.715 \text{ cm}$$

Thus, the estimated linear regression equation is:

$\hat{y} = a + bx = \hat{y} = 0.715 + 0.27x$ for $3 \leq x \leq 17$ (avoid extrapolating the regression line beyond the range of observations)

This is the estimated linear functional relationship between age (days) and wing length (cm). Thus, wing length increases by 0.27 cm every day.

Step 3: Test the significance of β (linear regression coefficient)

To test β , compute the residual mean square as:

$$S^2_{yx} = \frac{\sum y^2 - \frac{(\sum xy)^2}{\sum x^2}}{n - 2} = \left(19.66 - \frac{(70.80)^2}{262.00} \right) / 13 - 2 = 0.05$$

The residual mean square denotes the variance of y after taking into account the dependence of y on x.

Compute the test statistic (tb) value as:

$$t_b = \frac{b}{\sqrt{\frac{S^2_{yx}}{\sum x^2}}} = \frac{0.27}{\sqrt{\frac{0.05}{262}}} = \frac{0.27}{0.0135} = 19.5$$

Step 4: Compare the calculated tb value with tabulated t-value at $\alpha/2$ level of significance, at n-2 (13-2) = 11 d.f.; where n is pair of observations.

At 5% level, $t_{0.025(11)} = 2.201$, and at 1% = $t_{0.005(11)} = 3.106$

Since calculated /tb/ (19.5) is > the tabulated t-value at the 1% level of significance, the linear response of wing length to changes in the days within the range of 3 to 17 days is highly significant.

12.2.2 Simple linear correlation analysis

The simple linear correlation analysis deals with the estimation and test of significance of the simple linear correlation coefficient (r), which is a measure of the degree of linear association between two variables x and y (there is no need to have a dependent and independent variable).

The value of r lies within the range of -1 to +1, with extreme values indicating the perfect linear association and the mid-value of 0 indicates no-linear association between the two variables. The value of r is negative when a positive change in one variable is associated with a negative change in another and positive when the values of two variables change in the same direction (increase or decrease).

Even though the zero r value indicates the absence of linear association between two variables, it does not indicate the absence of association between them. It is possible for the two variables to have a non-linear association such as quadratic form. The procedure for the estimation and test of significance of a simple linear correlation coefficient between two variables x and y are:

Step 1: Compute the means (\bar{x}, \bar{y}), the sum of square of the deviates ($\sum x^2$, and $\sum y^2$), and the sum of the cross product of deviates ($\sum xy$) of the two variables.

Step 2: Compute the simple linear correlation coefficient for the above example as:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{70.8}{\sqrt{262 \times 19.66}} = \frac{70.80}{71.77} = 0.98$$

Step 3: Test the significance of the simple linear correlation coefficient (r) by comparing the computed r -value with the tabulated r -value at $n-2$ d.f. The simple linear correlation coefficient (r) is declared significant at α level of significance if the absolute value of the computed r -value $>$ the corresponding tabulated r -value.

Computed $r = 0.98$ with d.f. of $n-2 = 13-2 = 11$
 r – table at 5% (11) = 0.55, r at 1% (11) = 0.68

Thus, the simple linear correlation coefficient is significant at 1% level of significance which indicates the presence of a highly significant and positive linear association between ages and wing length of sparrows.

12.3 Multiple Linear Regression

The simple linear regression and correlation analysis is applicable only in cases with one independent variable. However, in many situations Y may be dependent on more than one independent variables. Linear regression analysis involving more than one independent variables is called multiple linear regression. The relationship of the dependent variable Y to the K independent variables X_1, X_2, \dots, X_k can be expressed as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

The data required for the application of multiple linear regression analysis involving K independent variables are $(n)(k + 1)$ observations, where n is number of pairs ($n > 2$).

Linear regression involving two independent variables can be expressed as: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ where β_1 & β_2 are partial regression coefficients. β_1 measures a change in Y for unit change in X_1 , if X_2 is held constant. Similarly, β_2 measures the rate of change in Y for a unit change in X_2 where X_1 is held constant. α (sometimes designated as β_0) is the value of Y when both X_1 & X_2 are zero.

For applying multiple linear regression analysis:

- the effect of each of K independent variables on Y must be linear
- the effect of each X_i on Y is independent of the other X_j , *i.e.* no interaction.

Example: The following data show the weight gain, initial body weight & age of five chicks fed with certain type of rations for a month.

Initial age (days) (X1)	5	5	5	4	6
Initial body weight (g) (X2)	10	15	12	15	20
Weight gain (g) (Y)	5	6	7	8	6

Fit the multiple linear regression model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ to the data

Step 1. Calculate the means and square of the deviates

	X1	X2	Y	y^2	x_1^2	x_2^2	$x_1 y$	$x_2 y$	$x_1 x_2$
	5	10	5	1.96	0	19.36	0	6.16	0
	5	15	6	0.16	0	0.36	0	-0.24	0
	5	12	7	0.36	0	5.76	0	-1.44	0
	4	15	8	2.56	1	0.36	-1.6	0.96	-0.6
	6	20	6	0.16	1	31.36	-0.4	-2.24	5.6
Sum	25	72	32	5.2	2	57.2	-2	3.2	5
Mean	5	14.4	6.4						

Step 2. Solve for b_1 & b_2

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(57.2 \times -2) - (5 \times 3.2)}{(2 \times 57.2) - (5)^2} = \frac{-130.4}{89.4} = -1.46$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(2 \times 3.2) - (5 \times -2)}{(2 \times 57.2) - (5)^2} = \frac{16.4}{89.4} = 0.18$$

Step 3. Compute the estimate of the intercept as:

$$a = \bar{Y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$a = 6.4 - (-1.46 \times 5) - 0.18 \times 14.4 = 6.4 + 7.3 - 2.59 = 11.11$$

Thus, the estimated multiple linear regression equation for initial age (days) and initial body weight (g) with weight gain (g) is: $\hat{Y} = 11.1 - 1.46 X_1 + 0.18 X_2$ for $4 \leq X_1 \leq 6$; and $10 \leq X_2 \leq 20$.

Step 4: Compute:

The sum of squares due to regression (SSR) = $\sum_{i=1}^k (b_i)(\sum x_i y)$

$$SSR = b_1 \sum x_1 y + b_2 \sum x_2 y = (-1.46 \times -2) + (0.18 \times 3.2) = 3.496$$

$$\text{Residual (error) SS} = \sum y^2 - SSR = 5.2 - 3.496 = 1.704$$

$$\text{Coefficient of determination (R}^2\text{)} = \frac{SSR}{\sum y^2} = \frac{3.496}{5.2} = 0.67$$

R^2 measures the amount of explained variation of Y due to the independent variables. Thus, in the above example 67% of the total variation in weight gain (g) of chicks can be accounted for a linear function involving initial age (days) and initial body weight (g).

Step 5: Test the significance of R^2

$$\text{Compute F value as: } \frac{SSR / k}{SSE / (n - k - 1)} = \frac{3.496 / 2}{1.704 / (5 - 2 - 1)} = \frac{1.748}{0.852} = 2.05$$

where k is number of independent variables (2) and n is number of data pairs (5)

Read table F-value as $F_{(k, n-k-1)}$; $F_{(2, 2)} 5\% = 19.00$; $F_{(2, 2)} 1\% = 99.00$

Since the computed F-value (2.05) is less than the table F value at 5% (19.00) the estimated multiple linear regression $\hat{Y} = 11.1 - 1.46 X_1 + 0.18 X_2$ is not significant at the 5% level. Thus, the combined linear effect of initial age (days) and initial body weight (g) on weight gain (g) of chicks is not significant.

The larger the R^2 value, the more important the regression equation in characterizing Y. On the other hand, if the value of R^2 is low, even if the F-test is significant, the estimated linear regression equation may not be useful. For example an R^2 value of 0.26, even if significant indicates that only 26% of the total variation in the dependent variable (Y) is explained by the linear function of the independent variables considered.

13. DATA TRANSFORMATION

For valid applications of parametric analysis like ANOVA, t-test, etc certain basic assumptions must be met. If the data violate such assumptions transformation of data can be used.

Data transformation can also be used to reduce the CV

The appropriate type of data transformation to be used depends on the specific type of relationship between the variances and the means. Conduct the analysis using the transformed data, and in tables present transformed means in parenthesis alongside their back transformed values out of parenthesis.

The three commonly used data transformation methods are:

13.1 Logarithmic Transformation

More appropriate when the treatment effects are multiplicative rather than additive, then the logarithmic transformation of the data will exhibit additivity. Such conditions are generally found on count data such as number of insects per plot, number of eggs per plant, etc.

- $X' = \log(x + 1)$; where x is original data and $x + 1$ is preferred especially when some of the observed values are small. Logarithmic of base 10 are generally used but any base would be satisfactory.
- Number of eggs/plant = 9; $\log(9+1)$; $\log(10) = 1$

13.2. Square Root Transformation

The square root transformation is applicable when the data consists of counts of rare events such as the number of infested plants in a plot, the number of insects caught in traps. For such data the variance tends to be proportional to the mean. Square root transformation is also appropriate for percentage data where the range is between 0-30% or between 70-100%.

If most of the values in the data set are small especially with zeros present.

$$X' = \sqrt{x + 0.5}; x = 0; \sqrt{0 + 0.5} = 0.707$$

Statistical computation can be done on the transformed data. The mean can be expressed in terms of the original data by squaring the transformed value and subtractions of 0.5

13.3 Arcsine Transformation

An Arcsine or angular transformation is appropriate for data on proportions and data expressed in decimal fractions or percentages. The arcsine transformation abbreviated “arcsine” frequently referred as “angular transformation” or “inverse sine” or “ \sin^{-1} ”. However, not all percentage data need to be transformed and Arcsine is not the only transformation possible.

- Rule 1: For percentage data lying within the range of 30 to 70%, no transformation is needed.
- Rule 2: For percentage data lying within the range of 0-30 or 70 – 100%, but not both, use the square root transformation.
- Rule 3: For percentage data that do not follow the ranges specified in either rule 1 or rule 2, the arcsine transformation should be used.

$$P' = \sin^{-1} \sqrt{P}, P = 100 \% = \sin^{-1} \sqrt{1} = 90^\circ. \text{ An angle whose sin is } 1 = 90^\circ.$$

For proportion of 0 to 1 (0-100%), the transformed values will range between 0 and 90 degrees, percentage 44%: $\sin^{-1} \sqrt{0.44} = 41.55^\circ$

Transformed values can be transformed back to proportion as:

$$P = (\sin P')^2 \quad P = (\sin 90)^\circ = 1 = 100\%, P = (\sin 41.55)^\circ = 0.44 \text{ or } 44\%$$