

CHAPTER 1

INTRODUCTION TO STATISTICS

Biometry is defined by combining two words such that; Bio (life) and metry (measurement). Therefore it is defined as the application of statistical methods and principles to the solution and analysis of biological problems. Biometry always associated with research and experiment. The biological problems are those arising in the basic biological sciences as well as in applied areas as the health-related and the agricultural sciences. Biometry is also called biological statistics or Biostatistics. Statistics in its modern sense defined as the scientific study of numerical data based on natural phenomena.

Scientific study: Statistics must meet the commonly accepted criteria of validity of scientific evidence.

Research: systematic approach to solve problems to investigate new facts or to verify the earlier result/hypothesis. The procedures we follow in the research are called research methodology. Hence, biometry is the main part of the research methodology in experiment.

1.1. Basic terminology and definition in Biometrics

Data: Statistics generally deals with populations or groups of individuals hence it deals with *quantities* of information, not with a single *datum*. Thus, the measurement of a single animal or the response from a single biochemical test will generally not be of interest but data on many records or results. Data can be qualitative or quantitative information taken on a certain character. For example, data of height, weight, color, etc

Distribution: The spread of statistics within known or possible limits, especially in relation to the norm or to expectations (i.e. the dispersion of a specific data for a given characteristics).

Normal distribution: The normal distribution is used to model some continuous variables. It is a symmetrical bell shaped curve that is completely determined by two parameters. They are the distribution (or population) mean, μ , and the standard deviation, σ . Such that between $(\mu - \sigma)$ and $(\mu + \sigma)$, $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$.

Hypothesis: A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

- Simply, it is a proposed explanation for a phenomenon.
- Or it can be defined as a tentative statement about the relationship between two or more variables.

Null hypothesis (H₀): No significant difference in a specific population with specific characteristics. It represents a theory that has been put forward, usually as a basis for argument.

Alternative Hypothesis (H_A or H_I): Is the opposite idea for “H₀”. In hypothesis testing a H₀ (typically that there is no effect) is compared with an H_I (typically that there is an effect, or that there is an effect of a particular sign).

- **For example** if comparing average lactation milk yield from improved feed and normal grazing. Then we could have: H_0 , that the two means are equal, i.e. there is no difference between the two types of milk yield. Thus the conclusion is given in terms of the null hypothesis. We either “Reject H_0 in favour of H_1 ” or “Do not reject H_0 ”. If we reject H_0 , then we declare the result to be “statistically significant”, and this provides evidence that H_1 is true. If we conclude “Do not reject H_0 ”, then the result is declared to be “not statistically significant”. This does not necessarily mean that H_0 is true, only that we do not have sufficient evidence to reject it.

Type I error: The incorrect rejection of a true H_0 (i.e. rejecting H_0 when it is true).

Type II error: Incorrect retaining (accepting) a false H_0 (i.e. accepting H_0 when it is false).

- Simply, a **type I** error is the false detection of an effect that is not present, while **type II** error is the failure to detect an effect that is actually present.
- **Type I** error is the probability of over reacting , while **type II** error is the probability of under reacting which are symbolized by the **Greek (α)** letter **alpha** and **beta (β)** respectively.
- **E.g.** - **Type I error** = Alarm with no fire
- **Type II error** = Fire with no alarm

P-value: The probability value (p-value) of a hypothesis test is the probability of getting a value of the test statistic as extreme, or more extreme, than the one observed, if the null hypothesis is true. Small p-values suggest the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the evidence to reject the null hypothesis. It was common to select a particular p-value, (often 0.05 or 5%) and reject H_0 if (and only if) the calculated probability was less than this fixed value.

Significance level, of a hypothesis test: The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0 , if it is in fact true. It is the probability of a type I error and is set by the investigator in relation to the consequences of such an error.

Confidence level (interval): The probability that the value of a parameter falls within a specified range of values. Confidence interval is a range of values that is likely to contain an unknown population parameter. So, if significance level is **0.05**, the corresponding confidence level is **95%**

Mean: The mean is a measure of the “middle”, “average” mostly with the symbol \bar{x} . As a formula, $\bar{x} = \sum x / n$. Where \sum is short for “the sum of”, “x” signifies that each value is taken in turn, and n is the number of observations.

Median: The median is the "middle value" of a list. In an odd number of entries, the median is the middle entry after sorting. But for even number of entries, the median is halfway between the two middle numbers after sorting.

Model: A combination of factors/functions that determine the values of certain variable.

Model (statistical): A statistical model is a simple description of a process that may have given rise to observed data.

Outlier: An outlier is an observation that is very different to other observations in a set of data. Since the most common cause is recording error, it is sensible to search for outliers (by means of summary statistics and plots of the data) before conducting any detailed statistical modelling.

Chi-square Statistic: The chi-square statistic is used to measure the agreement between categorical data and a multinomial (different character) that predicts the relative frequency of outcomes in each possible category

Population: A population is a collection of units being studied. The biological definition of population refers to all the individuals of a given species (perhaps of a given life-history stage or sex) found in a circumscribed area at a given time.

Sample: A sample is a group of units, selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions (inferences) about the population. A sample is usually used because the population is too large to study in its entirety. The sample should be representative of the population. This is best achieved by random sampling.

Random sample: Taking of sample with unbiased manner.

Precision: Precision is a measure of how close an estimator is expected to be the true value of a parameter. Precision is usually expressed in terms of the standard error of the estimator. Less precision is reflected by a larger standard error.

Range: The range is the difference between the maximum and the minimum values. It is a simple measure of the spread of the data.

Confounding: When the differences between the treatment and control groups other than the treatment produce differences in response that are not distinguishable from the effect of the treatment, those differences between the groups are said to be *confounded* with the effect of the treatment (if any). Age

Controlled experiment: An experiment that uses the methods of comparison to evaluate the effect of a treatment by comparing treated subjects with a control group, who do not receive the treatment.

Correlation: A measure of linear association between two (ordered) lists. Two variables can be strongly correlated without having any causal relationship, and two variables can have a causal relationship and yet be uncorrelated.

Correlation coefficient: The correlation coefficient r is a measure of how nearly a scatter plot falls on a straight line. The correlation coefficient is always between -1 and $+1$. To compute the correlation coefficient of a list of pairs of measurements (X,Y).

Linear regression: is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variable (or independent variables) denoted X .

Cross-sectional study: A cross-sectional study compares different individuals to each other at the same time.

Skew (skewness): If the distribution (or “shape”) of a variable is not symmetrical about the median or the mean it is said to be skew. The distribution has positive skewness if the tail of high values is longer than the tail of low values, and negative skewness if the reverse is true.

Kurtosis: Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers

Spread, measures of: Most data sets exhibit variability -- all values are not the same! Two important aspects of the distribution of values are particularly important; they are the centre, and the spread. The “centre” is a typical value around which the data are located. The mean and median are examples of typical values. The spread describes the distance of the individual values from the centre.

The range (maximum – minimum) and the inter-quartile range (upper quartile – lower quartile) are two summary measures of the spread of the data. The standard deviation is another summary measure of spread.

Standard deviation: The standard deviation (S.d) is a commonly used summary measure of variation or spread of a set of data. It is a “typical” distance from the mean.

Standard error: The standard error (s.e) is a measure of precision. It is a key component of statistical inference. The standard error of an estimator is a measure of how close it is likely to be, to the parameter it is estimating.

Transforming variables: If there is evidence of marked skewness in a variable, then applying a transformation may make the resulting transformed variable more symmetrical.

Variable or character: The *actual property* measured by the individual observations. The more common term employed in general statistics is "variable." The characteristic measured or observed when an observation is made. Variables may be non-numerical (categorical or factor variable) or numerical. However, in biology the word "character" is frequently used synonymously.

Independent Variable: In regression, the independent variable is the one that is supposed to explain the other; the term is a synonym for "explanatory variable." Usually, one regresses the "dependent variable" on the "independent variable."

Dependent variables: The dependent variable is what is being studied and measured in the experiment. It's what changes as a result of the changes to the independent variable. An example of a dependent variable is how tall you are at different ages. The dependent variable (height) depends on the independent variable (age).

Mixed variable: Some variables are between being categorical and numerical. For example daily rainfall is exactly zero on all dry days, but is a continuous variable on rainy days. Wind speed is similar, with zero being calm.

Ordinal variable: An ordinal variable is a categorical variable in which the categories have an obvious order, e.g. (strongly disagree, disagree, neutral, agree, strongly agree).

Variability or variation or dispersion: The variability (or variation) in data is the extent to which successive values are different.

Variance: The variance is a measure of variability, and is often denoted by s^2 . In simple statistical methods the square root of the variance, s , which is called the standard deviation, is often used more.

Treatment: The substance or procedure studied in an experiment or observational study. At issue is whether the treatment has an effect on the outcome or variable of interest.

Treatment Effect: The effect of the treatment on the variable of interest. Establishing whether the treatment has an effect is the point of an experiment.

Treatment group: The individuals who receive the treatment, as opposed to those in the control group, who do not.

1.2. Definition and meaning of Design of Experiments

Experimental design (design of experiments) is the process of planning a study to meet a specified objective. An experiment can be defined as planned research conducted to obtain new facts, or to confirm or refute the results of previous experiments. Most generally, observing, collecting or measuring data can be considered an experiment. In a narrow sense, an experiment is conducted in a controlled environment in order to study the effects of one or more categorical or continuous variables on observations.

Planning a study:

1. Survey: Statistical decision based on raw data and own observation
2. Experimentation: Decision based on carrying experiment using experimental units with appropriate design.

The scientific method:

The "scientific method" is a formal statement of procedure designed to facilitate the scientist's making the most effective use of his or her observations. The scientific method is usually defined to consist of the following four steps.

1. **Formulation of the hypothesis:** Based on preliminary observations, this is the tentative explanation.
2. **Planning the experiment:** The experiment must be constructed to objectively test the hypothesis. This is what this course is all about.
3. **Careful observation and collection of the data:**
4. **Interpretation of the results:** The results of the experiment may lead to confirmation, alteration, or rejection of the hypothesis.

Some important characteristics of a well-planned experiment are:

1. **Degree of precision:** The probability should be high 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.
2. **Simplicity:** The design should be as simple as possible, consistent with the objectives of the experiment.

3. **Absence of systematic error:** Experimental units receiving one treatment should not differ in any systematic way from those receiving another treatment so that an unbiased estimate of each treatment effect can be obtained.
4. **Range of validity of conclusions:** Conclusions should have as wide a 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 of increasing the range of validity of an experiment.
5. **Calculation of degree of uncertainty:** The experiment should be designed so that it is possible to calculate the possibility of obtaining the observed result by chance alone.

Steps in experimentation:

An experiment is usually planned and can be described in several steps.

1. *Problem(s) identification*
2. *Statement of the Hypothesis(es)*
3. *Determine the objectives*
4. *Select the treatments and the experimental material*
5. *Select the experimental design*
6. *Select the experimental unit and number of replications*
7. *Ensure proper randomization and layout*
8. *Ensure proper means of data collection*
9. *Outline the statistical analysis before doing the experiment*
10. *Conduct the experiment*
11. *Analyze the data and interpret the results relative to the hypothesis*
12. *Prepare complete and readable reports (give a conclusion)*

The planning of an experiment begins with an introduction in which a problem is generally stated, and a review of the relevant literature including previous results and a statement of the importance of solution of the problem.

Definition of Terminologies in Experimental Design

- ➡ **An experimental study:** is a scientific test that investigates the relationship between an outcome and one or more conditions manipulated by the researcher.
- ➡ **An experiment:** is “deliberate observation under conditions deliberately arranged by the observer”.
 - Before considering the appropriate experimental design, it is important to be clear about the aims of any experiment, which are usually associated with one or more scientific questions or hypotheses to be tested.
 - A thorough definition of the objectives of the design is required to make easy to assess whether the chosen treatments are sufficient to assess these aims.

- ➡ **Treatment:** is the set of different experimental conditions to be tested. A treatment is any manipulation that a researcher can perform.
 - Asides from identifying the experimental treatments, **experimental units** must be chosen.
 - Treatments are chosen to enable the experimental hypotheses to be tested.
- ➡ **Experimental Unit (EU):** Is “*the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment*”.
 - Experimental Units should be chosen to be **reasonably homogeneous** so that treatments are compared on a fair basis.
 - Experimental Units are chosen according to the frame of reference for the experiment.
- ➡ **Measurement units:** are units on which the individual observations are recorded.
- ➡ **An observation:** is a measurement taken upon an experimental unit.
- ➡ **Variable:** A measurable characteristic of a plot
- ➡ **Observation (also variate):** An individual measurement of a variable
- ➡ **Population:** The set of all possible values of a variable.
- ➡ **Parameter:** A fixed value that characterizes a population
- ➡ **Sample:** A set of measurements that constitutes part of a population
- ➡ **Statistic:** The value of a parameter as applied to a sample
- ➡ **Controls:** are objects that receive a null or neutral treatment.

Natural variation may be inflated/overstated by **measurement variation** (also known as measurement error). This combined background variation is a potential cause of both bias and uncertainty in experimental results.

Aims of experimental design and statistical analysis:

- ✓ Distinguish the performance, character or some ones identity
- ✓ Quantify some characters and
- ✓ Compare the effect of treatments (signal) and background variation (noise).

Experimental bias is avoided by *random allocation* of treatments to EU (i.e. *randomization*).

Precision is attained by proper *replication* and *blocking*.

Specific issues of experimental design:

Once the objectives, interesting questions, and the hypothesis are defined, the scope, type, and requirements of an experiment are also more or less determined. Thus, *the experiment should be designed to meet those requirements*.

Specifically, experimental design is concerned with the following issues:

1. **The size of the study:** number of replications, and the size and shape of experimental units.
2. **Type and number of measurements:** availability of a measuring device, precision and accuracy of the measurement, and the timing of making measurements.
3. **Treatments:** the type of treatments, the levels of treatment, and the number of treatments.

4. **Assignment of treatments to experimental units:** completely random, restricted randomization, etc.
5. **Error control:** the error control can be accomplished by blocking techniques, the use of concomitant observations, the choice of size and shape of the experimental units, and the control of the environment using a growth chamber or greenhouse.
6. **Relative precision of designs involving few treatments:** to compare two experimental designs one compares amounts of information.

Why We Use Experimental Designs?

- a. Experimental designs are used so that the treatments may be assigned in an organized manner to allow valid statistical analysis to be carried out on the resulting data.
- b. Different designs isolate different known or suspected sources of variation so that the treatments effects can be evaluated free of extraneous environmental or other influences.
- c. Statistical theory also requires that certain conditions be met during the execution of the experiment to permit valid probability statements to be made.

Purposes of experimental designs:

- a. To provide estimates of a treatment effects or differences among treatment effects.
- b. To provide an efficient way of confirming or denying 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 variation from the comparisons of interest.
- f. To provide a systematic, and efficient pattern of conducting an experiment.

1.3. Types and Principles of Agricultural Experimental design

Research can be classified based on different factors:

- ✚ Based on nature of observations: Absolute and Comparative Experiments
- ✚ Based on factors to be tested: One factor, two factor and multifactorial experiment
- ✚ Based on stage of experimentation: Confirmatory and exploratory experiments

There are three basic principles of any experimental design are replication, randomization and blocking.

1. Replication:

- Is the process of applying each treatment to more than one Experimental Unit (EU).
- The number of independent EU to which the treatment is applied to is the number of replicates.
- ✓ **Replication** refers to the number of experimental units that are treated alike (Experimental unit or experimental plot is the unit of material to which one application of a treatment is applied).

Uses of replication:

- ❖ Repeating each treatment on several EU attains a more *reliable* estimate of the effect of each treatment.
- ❖ Replicated observations provide an estimate of *background variation* between units, which can be used to assess the importance of treatment differences.
- ❖ To provide an estimate of the experimental error. The experimental error is the variation which exists among observations on experimental units treated alike.
 - When there is no method of estimating experimental error, there is no way to determine whether observed differences indicate real differences or are due to inherent variation.
- ❖ To improve the precision of an experiment by reducing the standard deviation of a treatment mean. Increased replication usually improves precision, decreasing the lengths of confidence intervals and increasing the power of statistical tests.
- ❖ To increase the scope of inference of the experiment by selection and appropriate use of more variable experimental units.
 - Example: replication in time and space in yield trials.
- ❖ **To effect control of the error variance.** The aim is to assign the total variation among experimental units so that it is maximized among groups and, simultaneously, minimized within. Experimental error must not be inflated by differences among groups.

Technical vs Biological replication

Technical replication involves the repeated measurement of the *same* sample. It is important in controlling for errors in measurement or technology. This type of replication is always considered pseudo-replication.

Biological Replication takes place when measurements are taken from several independent biological subjects rather than from a single individual. It is important in the statistical inference of populations.

2. Randomization:

Randomly allocating treatments to Experimental Unit, to ensure fair assessment of the treatments guarding the design against *bias* and coping with the *natural variation* between EUs.

Randomization is the assignment of treatments to experimental units so that all units considered have an equal chance of being assigned a given treatment within the study. It functions to assure unbiased estimates of treatment means and experimental error.

Methods for randomly allocating treatments to units

- ➡ Flip coin, throw a die, pick numbers from a hat, select cards from a pack, use random number tables and run computer packages

3. Blocking:

The process of identifying or building groups of EU which are expected to have similar responses in the absence of any treatment effects.

- ❖ *Blocks* are subsets of experimental material within which EU are expected to be homogeneous, with more heterogeneity allowed between EU in different blocks.
- ❖ It is not always possible to have reasonably homogeneous EUs (as they are essentially heterogeneous).
- ❖ However, it is possible to identify groups of EUs, such that within such groups EUs are *reasonably* homogeneous, but heterogeneous across different groups.
- ❖ Information on the causes of heterogeneity is used to define blocks.
- ❖ Treatments randomized to units within each block separately
- ❖ Variation is controlled and estimates of background noise reduced

CHAPTER 2

ANALYSIS OF VARIANCE AND COMMON EXPERIMENTAL DESIGNS

2.1. Analysis of variance (ANOVA)

- ➡ Analysis of variance is a technique for exploring the variation of a continuous response variable (dependent variable). The response variable is measured at different levels of one or more classification variables (independent variables).
- ➡ ANOVA is a statistical procedure for summarizing a classical linear model; a decomposition of sum of squares into a component for each source of variation in the model along with an associated test (the F-test) of the hypothesis that any given source of variation in the model is zero.
- ➡ ANOVA represents a set of models that can be fit to data, and also a set of methods that can be used to summarize an existing fitted model.
- ➡ The **ANOVA** is also statistical tool for splitting variability b/n and within group means in to component sources. These components can be thought of as the **signal (b/n treatment means)** and the **noise (within trt means/ error term)**. The signal is seen as differences among group means. The noise is seen as variability within groups. By measuring the variability within groups one has a baseline against which differences among group means can be compared.
- ➡ ANOVA is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken.
- ➡ ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that cause the mean in one group to differ from the mean in another.

The idea of the analysis of variance is to take a summary of the variability in all the observations and partition it into separate **sources**. The first component is total **sum of squares w/c** is partitioned into two separate, and additive, pieces. These are a **sum of squares among** (SS_{Among}) and a **sum of squares within**, (SS_{Within}). The SS_{Within} accumulates variability from within each group. While, SS_{Among} measures variability due to differences among the group means.

Note: $SS_{Among} + SS_{Within} = SS_{Total}$

- ➡ Associated with each sum of squares is a **degree of freedom**. In general, one starts with N degrees of freedom and loses one degree of freedom for **every sample mean** calculated. For the SS_{Total} there is one grand sample average, therefore there are $N - 1$ degrees of freedom.

There are $n_i - 1$ degrees of freedom within each group. Therefore, there are $\sum (n_i - 1) = N - k$ degrees of freedom for SS_{Within} . That leaves $k - 1$ degrees of freedom for SS_{Among} .

- ➡ Higher degrees of freedom generally mean larger sample sizes, and also a higher degree of freedom means more power to reject a false null hypothesis and find a significant result.
- ➡ In ANOVA the number of degrees of freedom can be interpreted as the number of independent comparisons that can be made among means in an experiment. Any constraint that the means must satisfy removes one degree of freedom.

Degrees of freedom within = $df_W = N - k$, N = total number of observations and
K = number of treatments.

- ➡ A sum of squares divided by its associated degrees of freedom produces a **mean square**. The sums of squares, degrees of freedom, and mean squares are all summarized in an analysis of variance (ANOVA) table. The mean square within is often referred to as the **error mean square**. Within sample variability is attributed to random (sampling) error. This is the baseline against which differences among group means are compared. The **mean square among** contains some of this error variability but also variability due to differences among group means.
- ➡ The F -ratio = MS_{Among}/MS_{Within} serves as a measure of the statistical importance or **significance** of the **differences among the group means**.
 - Values of F close to one indicate that the differences among group means can be attributed to natural or random error variability.
 - Values of F much larger than one indicate that some of the groups differ significantly in terms of their mean or average values.
 - The cutoff between “close to one” and “much larger than one” can be found in a table of the F distribution. This tabulation assumes that the original observations are normally distributed with a common error variance.

Variance between Groups = $s_B^2 = \frac{SS_B}{df_B}$. This statistic, also called the **Means Square between (MSB)**, is a measure of the variability of group means around the grand mean.

Variance within Groups = The **variance within SW**) quantifies the spread of values within groups.

The variance within is also called the **Mean Square within (MSW)** and is calculated:

$$s_W^2 = \frac{SS_W}{df_W} \quad \text{Where,} \quad SS_W = \sum_{i=1}^k (n_i - 1) s_i^2$$

The ratio of the variance between and the variance within (s_W) is the ANOVA F statistic.

$$F_{stat} = \frac{s_B^2}{s_W^2}$$

Assumptions of ANOVA:

- (i) All populations involved follow a normal distribution.
- (ii) All populations have the same variance (or standard deviation).
- (iii) The samples are randomly selected and independent of one another.

Steps of Computing ANOVA and testing hypothesis of treatment effects

1. State the null and alternative hypotheses

- The null hypothesis for an ANOVA always assumes the population means are equal.

Hence, we may write the null hypothesis as:

$$H_0: \mu_3 = \mu_2 = \mu_1 \dots$$

- Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal.

Thus, the alternative hypothesis (H_1) is:

$$H_1: \text{At least one mean is not statistically equal, } \mu_3 \neq \mu_2 \neq \mu_1 \dots$$

2. Calculate the appropriate test statistic, F-statistic

The test statistic in ANOVA is the ratio of the between and within variation in the data. It follows an F distribution.

This includes computation of

- ✓ Total, between groups and within groups/error df,
- ✓ Total Sum of Squares (between and within sum squares of means/groups,
- ✓ B/n and within mean squares.

3. Obtain the Critical Value (F-tabulated value)

- ✓ To find the critical value from an F distribution you must know the numerator (MSTR) and denominator (MSE) degrees of freedom, along with the significance level.
- ✓ Important components are df of b/n groups, within groups and significance level ($\alpha = 1\%, 5\%, 10\% \dots$).

4. Decision Rule

You reject the null hypothesis if: **F (observed value) > F tabulated/ critical value**. i.e you accept alternative hypothesis. **UNLESS** you accept H_0 IF: $F_{cal} \text{ value} < F_{tab} \text{ value}$, i.e you reject the null hypothesis.

5. Interpretation

- ✓ If we rejected the null hypothesis, we are 95% confident (1- α) that the means are not statistically equal for d/t treatments.
- ✓ However, since only one mean must be different to reject the null, we do not yet know which mean(s) is/are different. In short, an ANOVA test will test us that at least one mean is different, but an additional test must be conducted to determine which mean(s) is/are different.
- ✓ **If you fail to reject the null hypothesis in an ANOVA then you are done.** You know, with some level of confidence, that the treatment means are statistically equal. However, if you reject the null then you must conduct a separate test to determine which mean(s) is/are different.
- ✓ There are several techniques for testing the differences between means, but the most common test is the Least Significant Difference Test.
- ✓ Least Significant Difference (LSD) for a balanced sample:

$$LSD = t_{\alpha/2, error\ df} \sqrt{\frac{2\ MS}{r}}$$

Where; MSE is the mean square error and r is the number of each treatment replicated

- ✓ Thus, if the **absolute value** of the difference between any two treatment means is greater than the LSD value, we may conclude that they are not statistically equal.
e.g /mean of trt 1 – mean of trt 2/ > LSD value, we conclude that there is significant d/c b/n trt 1 and 2 and so on.

Comparisons among means: Planned and unplanned comparisons:

We usually complete an ANOVA of more than two groups by examining the data in greater detail, testing which means are different from which other ones or which groups of means are different from other such groups or from single means. If there was no significant difference between the two groups no further test is needed. But actually there may be a difference b/n groups that is not significant. Even if there had been such a difference, no further tests are possible.

Planned tests are designed and chosen independently of the results of the experiment. They should be planned **before** the experiment has been carried out and the results obtained. Such comparisons are called planned or a **priori comparisons**. Such tests are applied regardless of the results of the **preliminary** overall ANOVA. E.g we compare the test treatments vs control groups. In this case we prior know the significant effect of the control.

By contrast, after the experiment has been carried out, we might wish to compare certain means that we notice to be markedly different. We might therefore wish to test whether there is in fact a significant difference between the effects of test treatments. Such comparisons, which suggest themselves as a result of the completed experiment, are called **unplanned** or a **posteriori comparisons**. These tests are performed only if the preliminary overall ANOVA is significant.

They include tests of the comparisons between all possible pairs of means. e.g we compare the actual difference b/n test treatments based on statistical significance.

Compute Coefficient of variation = it measures proportion of variability b/n treatments.

$$CV = \sqrt{\frac{EMS}{GM}} \times 100 \quad \text{or} \quad CV = (sd/mean) * 100\%$$

Types of analysis of variance (ANOVA)

a. One-Way ANOVA

- ❖ The One-Way ANOVA task **compares** the means of the response variable over the groups defined by a single classification variable.
- ❖ One-way ANOVA is an extension of the t-test to three or more samples
- ❖ It focus analysis on group difference of treatment means
- ❖ The One-Way ANOVA **used to** perform an analysis of variance when there is a continuous dependent variable and a single classification variable.
- ❖ The analysis of variance performed in the One-Way ANOVA task **indicates** whether the means of the groups are different; it does not indicate which particular means are different. E.g. CRD, RCBD (if there is no interaction b/n factors)

b. Two-Way ANOVA

- ❖ Two-way ANOVA (and higher) focuses on the interaction of Factors
- ❖ It focuses on the research question “Does the effect due to one factor change as the level of another factor changes?” eg. RCBD, LSD, Split-plot design.

c. Factorial ANOVA: The Factorial ANOVA task enables you to perform an analysis of variance when you have multiple classification variables. A factorial model is specified, and a plot of the two-way effects is requested.

d. Linear and mixed Model ANOVA:

- ✓ The Linear Models task enables to compare means and explain variation when we have a model that includes classification variables, quantitative variables, or both (such as in an analysis of covariance).
- ✓ This linear model and additionally requests a retrospective power analysis and a plot of the observed values versus the predicted values.
- ✓ The Mixed Models task enables to fit basic mixed models. A mixed model is a linear model that contains both fixed effects and random effects.
e.g crossover design, repeated design, ... etc

e. Multivariate ANOVA

Is used when there are more than one dependent and several independent variables. **e.g** effect of parity, age, season and nutrition on milk yield or butter fat percentage in dairy cows.

- ➡ Dependent variables: Milk yield and butter fat%
- ➡ Independent variables: parity, age, season, nutrition,... etc

Assumptions of any ANOVA:

The theoretical concepts of ANOVA are based on a set of assumptions:

1. All ANOVAs require that sampling of individuals be at random. It is adequate safeguard to ensure random sampling during the design of experiment, or when sampling from natural population is required.
2. The items are being independent: It may be not true when some variables are correlated
3. Equality of variances: it is important to check that homogeneity of variance, **Homoscedasticity**)
4. The residual effect or random error is normally distributed, **normality**
5. The additive effect in two-way or higher ANOVA should be small.

2.2. Common experimental designs

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

Design of Experiments:

There are many types of experimental designs. They can be broadly classified as a single factor and multi-factor (factorial) experiments.

Most commonly used agricultural research designs are:

- Completely Randomized Design (CRD)
- Randomized Complete Block Design (RCBD)
- Latin Square Design (LSD)
- Split- Plot Design

However, the procedure for all research design is generally known as the scientific method, which, although difficult to define precisely. 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

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). In the study of the effect of different rations on milk production, ration is a treatment and animal is the experimental unit; while in the 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.

Therefore, every possible effort should be made to reduce the experimental error.

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.
- d) **Replication:** A situation where a treatment appears more than once in an experiment, it is said to be replicated.
- e) **Randomization:** Assigning the treatments to the experimental units in such a way 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.

2.2.1. Completely Randomized Designs (CRD)

When the treatments consists different levels of a single variable factor and all other factors kept at a single prescribed level, it is known as a single factor experiment. CRD is the basic single factor experiment. All other designs like RCBD and LSD stem from it by improving restrictions upon the allocation of the treatments within the experimental material. CRD are the simplest design in which the treatments are assigned to the experimental units completely at random. This allows every experimental unit, i.e., plot, animal, soil sample, etc., to have an equal probability of receiving a treatment.

Example: The 4 replicates of the 4 treatments are assigned at random to the 16 experimental units. This may be done using a table of random numbers, or by pulling numbered slips out of a hat. The analysis of variance table is also shown for this design.

Advantages of completely randomized designs

1. Complete flexibility is allowed - any number of treatments and replicates may be used.
2. Relatively easy statistical analysis, even with variable replicates and variable experimental errors for different treatments.
3. Analysis remains simple even data are missing.
4. Provides the maximum number of degrees of freedom for error for a given number of experimental units and treatments.

Disadvantages of completely randomized designs

1. Relatively low accuracy due to lack of restrictions which allows environmental variation to enter experimental error.
2. Not suited for large numbers of treatments because a relatively large amount of experiment material is needed which increases the variation.

Appropriate use of completely randomized designs

1. Under conditions where the experimental unit is homogeneous, i.e., laboratory, or growth chamber experiments.
2. Where a fraction of the experimental units is likely to be destroyed or fail to respond.
3. In small experiments where there is a small number of degrees of freedom.

The completely randomized design is seldom used in field experiments where the randomized complete block design has been consistently more accurate since there are usually recognizable sources of environmental variation.

Layouts in research design

Placement of the treatments on the experimental units along with the arrangement of experimental units is known as the layout of the experiment. Suppose that there are t treatments on the experimental units, namely $T_1, T_2, T_3, \dots, T_t$. Further, suppose that treatments are replicated r times each. If practical limitations dictate unequal replication should be made. We require, $txr = N$ experimental units. In case of equal replications the number of experimental units required will be $r_1 + r_2 + r_3 + \dots + r_t = N$.

The entire experimental material is divided into “ n ” numbers of experimental units. The units are numbered serially starting with one end and proceeding in a serpentine manner. For example, suppose there are five treatments with four replications. We need 20 experimental units. The 20 units are numbered as follows. i.e treatment 1 is applied to units 18, 4, 10 and 9; treatment 2 is applied to units 14, 7, 19 and 13., and so on. The final layout will be as follows:

1	2	3	4	5
T4	T3	T4	T1	T3
10	9	8	7	6
T1	T1	T3	T2	T3
11	12	13	14	15
T5	T4	T2	T2	T4
20	19	18	17	16
T5	T2	T1	T5	T5

Assumptions:

1. Independent observations
2. Normally distributed data for each group
3. Equal variances for each Trt groups

Mathematical Model

$$Y_{ij} = \mu + T_i + e_{ij}$$

Where Y_{ij} = an observation for which T_i (i and j denote the level of the factor and the replication within the level of the factor, respectively)

μ = is the average (mean) of all the data

T_i = is the effect of treatment level i

e_{ij} = the residual/ experimental error

Model ANOVA Table and variance partitioning

Source of variationn(SV)	df	SS	MS	Fcal	Ftab	
					5%	1%
Trt	t-1	*	*	*		
Error	t(r-1)	*	*			
Total	rt-1	*				

For paired comparisons the LSD value is computed as:-

$$\text{LSD} = t_{\alpha/2} \text{ SE (d)}$$

The Standard Error: It is the measure of the deviation of a mean of a sample from the means of other samples drawn from the same population. It is denoted by **S.E.** and is the ratio of **SD** to the root of the sample size (n). i.e.,

$$S.E. = S_{\bar{x}} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}} \quad S.E. = \frac{\text{Sample standard deviation}}{\sqrt{\text{number of observations}}} = \frac{s}{\sqrt{n}}$$

Examples

1. Hay was stored using three different methods (S1, S2 and S3) and its dry matter content was measured. Is there a significant difference among different storage methods?

S1	S2	S3
17.3	22.0	19.0
14.0	16.9	20.2
14.8	18.9	18.8
12.2	17.8	19.6

Solution

1. Define null and alternative hypotheses
2. Calculate grand total and grand mean, $GT = \text{Summation of all observations}$
Grand mean (GM) = GT/N
3. Calculate the adjustment/ correction factor
 $CF = GM^2 / rt$
4. Total sum square
 $\text{Squared Sum of each observation minus CF}$
5. Treatment sum of squares
 $\text{Summed square of treatments minus CF}$

6. Error sum of squares

$$\text{SS error} = \text{SS total} - \text{trt SS}$$

7. Mean of squares of treatment

$$\text{Trt SS} / \text{trt df}$$

8. Error mean squares

$$\text{Error SS} / \text{df}_{\text{error}}$$

9. Calculate F- statistic

$$\text{F- Value} = \text{MStrt} / \text{MSerror}$$

11. Compare the calculated F-value with critical value, and make decision

12. Calculate CV, coefficient of variability

$$\text{CV} = \sqrt{\frac{\text{EMS}}{\text{GM}}} \times 100\%$$

13. If it is significant at 5%, check which storage method results in higher DM content. Using least significant difference.

$$\text{LSD} = \sqrt{2\text{EMS}/r} \quad (t_{\alpha/2, \text{error df}})$$

15. Calculate confidence interval for trt means.

2.2.2. Randomized Complete Block Designs (RCBD)

The most frequently used exp'tal design is RCBD. RCBD differ from the CRD in that the experimental units are grouped into blocks according to known or suspected variation which is isolated by the blocks. The fundamental idea of the RCBD is to group experimental material together in to homogeneous blocks. Such a group is called a block or replication. Each treatment appears an equal number of times usually once, in each block and each block contains all the treatments. Therefore, the object of this grouping is to keep the errors within each group as small as possible.

Variation such as fertility, sex, slope, wind gradients, body weight, parity, age and litter size of animals can be isolated by appropriate blocking. Therefore, within each block, the conditions are as homogeneous as possible, but between blocks, large differences may exist. This results in relatively small gradients within each block so that the treatments may be compared under relatively homogeneous conditions. The treatments are assigned within the individual blocks at random with a separate randomization for each block. This reduces error term and more precise test of the treatment effects since the mean square for error will be smaller and the F value for treatment should be larger. RCBD is one of the most widely used designs. If it will control the variation in a particular experiment, there is no need to use a more complex design. The completely randomized design is seldom used in field experiments where the randomized complete block design has been consistently more accurate since there are usually recognizable sources of environmental variation.

Advantages of randomized complete block designs

1. **Flexibility:** Can have any number of treatments and blocks
2. Provides more accurate results than the completely randomized design due to grouping.
3. Relatively easy statistical analysis even with missing data.
4. Allows calculation of unbiased error for specific treatments.
5. If the variance is larger for some treatments than others, an unbiased error for testing any specific combination of the treatment means can be obtained.

Disadvantages of randomized complete block designs

1. Not suitable for large numbers of treatments because blocks become too large.
2. Not suitable when complete block contains considerable variability.
3. Interactions between block and treatment effects increase error.

Appropriate use of randomized complete block designs

1. When there is a known or suspected source of variation in one direction.
2. Adjust the blocks to have minimum variation within the block and orient plots to sample the entire range of variation within the block.
3. The randomized complete block design is one of the most widely used designs. If it will control the variation in a particular experiment, there is no need to use a more complex design.

An experiment with 4 treatments (A, B, C, D) and 3 blocks

1 A	1B	1D	1C	Block 1
2C	2 D	2B	2A	Block 2
3 B	3C	3A	3D	Block 3

ANOVA table of RCBD

SV	DF	SS	MS	Fcal	Ftab.
Trt	t-1	$\sum Ti/r$	SS/dft	TMS/EMs	
Blk	r-1	$\sum Rj/t$	SSR/dfb	BMS/EMs	
Error	(t-1)(r-1)	By sub.	ESS/Edf	-	
Total	rt-1	$\sum TR/rt$	-	-	

Trt = treatment, DF = Degree of freedom, SV = Source of variation, SS = Sum square, MS = Mean square, Fcal = Fcalculated, Ftab = Ftabulated

Decision: reject H_0 , if $F_{cal} > F_{tab}$ and accept H_0 if $F_{cal} < F_{tab}$ at specified significance level.

Randomization and Layout of RCBD

Step 1: Divide the experimental area (unit) into r-equal blocks, where r is the number of replications, following the blocking technique.

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

Step 3: Repeat step 2 for each of the remaining blocks. The major difference between CRD and RBD is that; randomization in CRD is done without any restriction to all experimental units but for RCBD, all treatments must appear in each block and different randomization is done for each block (randomization is done within blocks).

Steps for preparing layout of RCBD:

Preliminary activities:

1. Choose the number of blocks (minimum 2); the number of blocks is the number of replications.
2. Choose treatments (assign random numbers or letters for each trt)
3. Treatments are assigned at random within blocks of adjacent subjects, each treatment once per block, i.e. any treatment can be adjacent to any other treatment, but not to the same treatment within the block.

➤ Randomization for both blocks and treatments is important

Layout of CRBD with 4 treatments (T₁, T₂, T₃, T₄), 3 blocks and 3 replications per trt

1 T ₁	2 T ₂	3 T ₃	4 T ₄	Block 1
24 T ₄	23 T ₃	22 T ₂	21 T ₁	Block 2
25 T ₂	26 T ₁	27 T ₃	28 T ₄	Block 3

The statistical model for RCBD:

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

Where,

Y_{ij} = the observation on the jth block and the ith treatment

μ = Population Mean

τ_i = Effect of treatment, i

β_j = Effect of block, j

ε_{ij} = Experiment error for treatments i in block j

Step 1: Arrange the data by treatments, blocks and replication.

Step 2: Outline the analysis of variance

Step 3: Compute Correction Factor (C.F.) = $\frac{G^2}{rt} = \frac{(\text{Grandtotal})^2}{\# \text{ of experimental units}}$

➤ If trt is appeared more than ones in a block, **CF = G²/rtb**

Step 4: Compute sum of squares

- i. Total sum of square (SS_{TOT})
- ii. Treatment Sum of Squares (SS_{TRT})
- iii. Block Sum of Squares (SS_{BLK})
- iv. Error sum of squares (SS_{RES}) = SS_{TOT} – SS_{TRT} – SS_{RES}

Step 5: Compute the mean square for block, treatment and error by dividing each sum of squares by its corresponding d.f. (treatment, block and error).

$$\text{Block mean square (MSB)} = \frac{SSB}{r-1}$$

$$\text{Treatment mean square (MST)} = \frac{SST}{t-1}$$

$$\text{Error mean square (MSE)} = \frac{SSE}{(r-1)(t-1)}$$

Step 6: Compute the F-value for testing treatment and block differences.

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

Step 7: Read table F- and compare the computed F-value with tabular F-value and make decision.

- F-table for comparing block effects, use block d.f. as numerator and error d.f. as denominator F (trt df, error df) at 5% and at 1%.
- F-table for comparing treatment effects, use treatment d.f. as numerator and error d.f. as denominator F (trt df, error df) at 5% and at 1% .
- If the calculated F-value is greater than the tabular F-value for treatments at 1%, it means that there is real difference among treatment means.

Step 8: Compute the coefficient of variability (CV) and confidence interval (CI).

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

$$CI = Y_i \pm \frac{t\alpha}{2} \sqrt{\frac{MSE}{r}}$$

Step 9: Calculate least significance difference to compare the difference b/n treatment means.

$$LSD = \frac{t\alpha}{2} \sqrt{\frac{2MSE}{r}}$$

Efficiency of Blocking and relative efficiency (R.E):

Blocking maximizes the difference among blocks and reduces the difference among EUs of the same block 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 efficiently is:

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

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

By comparing it with tabular F-value (block df, error df) at 5% and 1%. If the computed F-value is greater than the tabular 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 < Table-F), it indicates either that the experimenter was not successful in reducing error variance by grouping of individual units or that the units were essentially homogeneous to start with.

Estimating Missing data in RCBD

Sometimes data for certain units are missing or become unusable.

For example,

- When an animal becomes sick or dies but not due to treatment.
- When rodents destroyed 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. An estimate of a missing value does not supply additional information to the experimenter; it only facilitates the analysis of the remaining data. When a single value is missing in RCBD, calculate an estimate of the missing value by:

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

Where:

Y= estimate of the missing data

t= # of treatments

r = # of replications

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

To = total of observed values in treatment containing the missing data.

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 make no contribution to the error sum of squares. When all of the missing values are on the same block or treatments the simplest solution is to act as if the block or treatment had not been included in the experiment.

EXAMPLE: In the table given below are dry matter yields (kg) of 4-varieties of forages (Alfalfa, Rhodes, sesbania, vetch) in 4-replications planted in RCBD for which one plot yield is missing. Estimate the missing value and analyze the data. Plot size = 10m x 10 m = 100m².

No	Treatments	I	II	III	IV	Treatment total (Ti)	Treatment mean
1	Alfalfa	18.5	15.7	16.2	14.1	64.5	16.1
2	Rhodes	11.7	Y	12.9	14.4	39(To)	12.9
3	Sesbania	15.4	16.6	15.5	20.3	67.8	16.7
4	Vetch	16.5	18.6	12.7	15.7	63.5	15.9
	Block total	62.1	50.9 (Bo)	57.3	64.5	234.8 = Go (Grand total)	

Go (Grand total) = **234.8**

Solution:

A. Estimate the missing value

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

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

B. Enter the estimated value and carry out the analysis:

- 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

$$1. \text{ CF} = \frac{(248.7)^2}{r \times t(4 \times 4)} = 3865.73$$

$$2. \text{ Total SS} = \sum_{i=1}^4 \sum_{j=1}^4 Y_{ij} - C.F. = (18.5)^2 + (11.7)^2 + \dots + (13.9)^2 + \dots + (15.7)^2 - C.F. = 79.18$$

$$3. \text{ Trt SS} = \frac{\sum_{i=1}^4 T_i^2}{r} - C.F. = \frac{(64.5)^2 + (52.9)^2 + (67.8)^2 + (63.5)^2}{4} - 3865.73 = 31.21.$$

$$4. \text{ Blocks SS} = \frac{\sum_{j=1}^4 B_j}{t} - C.F. = \frac{(62.1)^2 + (64.8)^2 + (57.3)^2 + (64.5)^2}{4} - 3865.73 = 9.02$$

$$5. \text{ Error SS} = \text{Total SS} - \text{treatment SS} - \text{block SS} = 79.18 - 31.21 - 9.02 = 38.95$$

D. Compute the correction factor for bias (B) for treatment sum of square

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

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

Y = estimated value;

$$B = \frac{[50.9 - (4-1)13.9]^2}{4(4-1)} = \frac{[50.9 - 41.7]^2}{12} = 7.05$$

E. Subtract the computed B value from total SS & treatment SS.

- Adjusted treatment SS = treatment SS – B
= 31.21 – 7.05 = 24.16
- Adjusted total SS = Total SS - B
= 79.18 – 7.05 = 72.13

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

Source of variation	Df	SS	MS	Computed F	Table F	
					5%	1%
Treatment	3	24.16	8.05	1.64		
Block	3	9.02	3.01	0.61		
Error	(t-1) (r-1)-1 = 8	38.95	4.9			
Total	rt-1-1 = 14	72.13				

$$CV = \frac{\sqrt{MSE}}{\text{Grand mean}} \times 100 = \frac{\sqrt{4.9}}{15.65(234.8/15)} \times 100 = 14.3\%$$

2.2.3. Latin Square Design (LSD)

Latin square design is used for a situation in which there are two extraneous sources of variation are blocked. 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 replication that is why it is called *Latin Square*. A Latin square is a table filled with “n” different treatments in such a way that each treatment occurs exactly once in each row and exactly once in each column.

Some basic principles:-

- Replicates are also included in this design.
- Treatments are assigned at random within rows and columns
- There are equal numbers of rows, columns, and treatments.
- Useful where the experimenter desires to control variation in two different directions

Squares of sizes 5 to 9 are generally used. For $t \leq 4$, there are too few error degrees of freedom. The opportunity to use large squares does not often arise. (Number of treatments is usually not very large).

A Latin square design is a method of placing treatments so that they appear in a balanced fashion within a square block or field. Treatments appear once in each row and column. Replicates are also included in this design.

- Treatments are assigned at random within rows and columns, with each treatment once per row and once per column.
- There are equal numbers of rows, columns, and treatments.
- Useful where the experimenter desires to control variation in two different directions.

A Latin square design is actually an extreme example of an incomplete block design, with any combination of levels involving the two blocking factors assigned to *one* treatment only, rather than to all.

The advantages of LSD:

- ➡ Greater precision is obtained than CRD & CRBD; because it is possible to estimate variation among row blocks as well as among column blocks and it is possible to remove them from the experimental error.
- ➡ They handle several factors and we either cannot combine them into a single factor or we wish to keep them separate.
- ➡ It allows experiments with a relatively small number of runs which are balanced.
- ➡ More precise than CRBD if no missed data exist.

The disadvantages are:

- ➡ The number of levels of each blocking variable **must equal** the number of levels of the treatment factor. So it needs balanced data.
- ➡ The Latin square model **doesn't assume interactions** between the blocking variables or between the treatment variable and the blocking variable.
- ➡ **Squares smaller than 5×5 are not practical** because of the small number of degrees of freedom for error.
- ➡ Latin square design is not appropriate design for large number of treatments. This is because one requires lots of replications which may not be attained, and also randomization procedure is difficult.

The Latin square design is for a situation in which there are two extraneous sources of variation. If the rows and columns of a square are thought of as levels of the two extraneous variables, then in a Latin square each treatment *appears exactly once in each row and column*.

The **mathematical model** for the classical Latin square design is

$$Y_{ijk} = \mu + Ti + Cj + Rk + e_{ijk}$$

Where, $i = 1, \dots, a$, $j = 1, \dots, a$, $k = 1, \dots, a$, is the observation for the experimental unit in the i^{th} row block level, j^{th} column block level and the k^{th} treatment effect.

Ti : effect due to treatment i

Cj : effect due to column j

Rk : effect due to row k

e_{ijk} Random error terms

Randomization and Layout:

To construct Latin square design for T number of treatments, we need T^2 number of experimental units. These units are classified into T groups of T units each based on one source of variation. This is called row classification. They will further be grouped into T groups of T units each based on the second source of variation. This is known as column classification. Because of additional blocking structure, there is some restriction in randomization. That is every treatment must occur only once in each column and row.

Step 1: To randomize five treatments in Latin Square Design, select a sample of 5×5 Latin square plan. Then you can create your 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:

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

Random numbers: 628 846 475 902 452

Rank: (3) (4) (2) (5) (1)

- Rank the selected random number from lowest to the highest.
- 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 plant (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:

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

ANOVA Table and partitioning of variance components:

There are four sources of variation in Latin Square Design, 2 more than that of CRD and one more than that for the RBD. The sources of variation are row, column, treatment and experimental error.

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

Example: Grain yield of three Elephant grass hybrids (A, B, and D) and a check variety, C, from an experiment with Latin Square Design. (Hint: variety = trts; row = soil type and column= fertilizer level). [t= 4, R= 4 and C=4]

<u>DM yield (t/ha)</u>					
Row #	Col 1	Col 2	Col3	Col 4	Row total (R)
Row 1	1.640(B)	1.210(D)	1.425(C)	1.345(A)	5.620
Row 2	1.475(C)	1.185(A)	1.400(D)	1.290(B)	5.350
Row 3	1.670(A)	0.710(C)	1.665(B)	1.180(D)	5.225
Row 4	1.565(D)	1.290(B)	1.655(A)	0.660(C)	5.17
Column total(C)	6.350	4.395	6.145	4.475	21.365
Grand total (G)					

Treatments totals:

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

Step 2: prepare the Outline of the analysis of variance table:

Source	D.F	SS	MS	Computed F	Table F	
					5%	1%
Treatment						
Row						
Column						
Error						
Total						

Compute source of variations and component values

Step 3. Compute the C.F. and the various Sum of Squares

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

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

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

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

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

$$\text{Error SS} = \text{Total SS} - \text{RoSS} - \text{Col SS} - \text{Treatment SS} = 1.41 - 0.03 - 0.83 - 0.43 = \mathbf{0.12}$$

Step 4. Compute the mean squares for each source of variation by dividing the sum of squares by its corresponding degrees of freedom.

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

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

$$\text{Treat. MS} = \frac{\text{Treatm. SS}}{t-1} = \frac{0.43}{3} = 0.143$$

$$\text{Error ms} = \frac{\text{Error SS}}{(t-1)(t-2)} = \frac{0.12}{3 \times 2} = 0.02$$

Step 5: Compute the F value for testing the treatment effect as: $F = \frac{TreatMS}{ErrorMS} = \frac{0.143}{0.02} = 7.15$

Step 6: compare the computed F value with tabulated F value.

Step 7: draw conclusion: Therefore, the compute F-value is higher than the tabular F value at 5% level of significance, but lower than the tabular F-value at the 1% level; the treatment difference is significant at the 5% level of significance.

Source	D.F	SS	MS	Computed F	Table F	
					5%	1%
Row	(t-1)=3	0.03	0.01	0.50 ^{ns}	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				

Compute the C.V.

$$C.V. = \frac{\sqrt{ErrorMS}}{G.mean} \times 100 = \frac{\sqrt{0.02}}{1.335} \times 100 = 10.6\%$$

Mean comparison:

Note that although the F-test on the analysis of variance indicates significant differences among the mean yields of the 4- grass varieties tested, *but it does not identify the specific pairs or groups of varieties that differed*. 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. It is the same as we used in case of RCBD.

$$LSD = t_{\alpha/2 \text{ error df}} \sqrt{\frac{2MSE}{r}} \quad \text{for equal replication, LSD= least significant difference}$$

$$LSD_{\alpha} = t_{\alpha/2 \text{ error df}} \sqrt{MSE \left(\frac{1}{r_i} + \frac{1}{r_j} \right)} \quad \text{for unequal replications}$$

Standard error of the difference between the treatment mean with a single missing value and a treatment mean with all units present.

$$s\bar{d} = \sqrt{MSE \left[\frac{2}{r} + \frac{1}{(r-1)(r-2)} \right]}$$

Effectiveness of Row and Column Blocking:

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 LS design indicate the gain in precision relative to either the CRD or RBD, the procedures are.

I. Compare the F-value for testing the row & column effects; and test their significance

$$F(\text{row}) = \frac{\text{Row } ms}{\text{Error } ms} = \frac{0.01}{0.02} = 0.50^{\text{ns}}$$

$$F(\text{column}) = \frac{0.275}{0.02} = 13.75^{**}$$

II. Compare the relative efficiency of LS design relative to CRD & RBD

a) The relative efficiency of LS design as compared to CRD

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

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

This indicates that the use of LS 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 time more replication would have been required to detect the treatment difference of the same magnitude as that detected with the LS design.

b) The R.E of LS design as compared to RBD can be computed in two ways:

- **When row is used as blocking factor**

$$\begin{aligned} \text{R.E (row)} &= \frac{Ms \text{ row} + (t-1)ms \text{ error}}{t(ms \text{ error})} \\ &= \frac{0.01 + (4-1) \times 0.02}{4 \times 0.02} = \frac{0.07}{0.08} = 0.875 \end{aligned}$$

- **When column used as blocking factor**

$$\begin{aligned} \text{R.E (column)} &= \frac{Ms \text{ column} + (t-1)ms \text{ error}}{t(ms \text{ error})} \\ &= \frac{0.275 + (4-1) \times 0.02}{4 \times 0.02} = \frac{0.335}{0.08} = 4.19 \end{aligned}$$

- 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 compared 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 used of Latin square design, is estimated to have increased the experimental precision over that of CRBD with rows as blocks by 290%, whereas the additional row-blocking in the LS design did not increase precision over the RBD with columns as blocks. Hence, for trial, a RBD with columns as blocks would have been as efficient as a LS design.

Missing Plots estimation in the Latin Square Design

The formula for single missing observations

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

Where R_o , C_o , and T_o are the totals of the observed values for the row, column, and treatment totals containing the missing value, and G_o is the grand total of the observed values. 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 RBD, the treatment sum of squares is biased upward by:

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

Where $[Go, Ro, Co \text{ and } T_o]$ are as described above. Then B is subtracted from treatment SS & total SS.

2.2.4. Split-Plot/strip Designs

What is a split-plot design?

In simple terms, a split-plot experiment is a blocked experiment, where the blocks themselves serve as experimental units for a subset of the factors. Thus, there are two levels of experimental units. The blocks are referred to as whole plots, while the experimental units within blocks are called **split Plots, split units, or subplots**.

Corresponding to the two levels of experimental units are two levels of randomization. One randomization is conducted to determine the assignment of **block-level** treatments to whole plots. Then, as always in a blocked experiment, a randomization of treatments to split-plot experimental units occurs within each block or whole plot.

Uses (importance) of split plot design:

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 subplots.

- 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.

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 subplot 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.

Advantages:

1. Experimental units which are large by necessity or design may be utilized to compare subsidiary (sub-plot trt) treatments.
2. Increased precision over a CRB design is attained on the subplot treatments and the interaction between subplot and main plot treatments.
3. The overall precision of the split plot design relative to the randomized complete block design may be increased by designing the main plot treatments in a Latin square design or in an incomplete Latin square design.
4. It permits the efficient use of some factors, which require large experimental units in combination with other factors, which require small experimental units.
5. It provides increased precision in comparison of some of the factors (sub –plot factors).
6. It promotes the introduction of new treatments into an experiment, which is already in progress.

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

Disadvantages:

1. The main plot treatments are measured with less precision than they are in a **RCBD**.
2. Low precision for the main plot factor can result in large differences being nonsignificant, while small differences on the sub-plot factor may be statically significant even though they are of no practical significance.
3. Statistical analysis is complicated because different factors have different error mean squares. And when missing data occur, the analysis is more complex than for a randomized complete block design with missing data.

4. Different treatment comparisons have different basic error variances which make the analysis more complex than with the randomized complete block design, especially if some unusual/ not critical type of comparison is being made.

Analysis procedure and ANOVA table

- ➡ Split plot design with 2 Main Plot Treatments (1, 2), 2 Sub Plot Treatments (A, B) & 4 Blocks

2A	2B	1B	1A	Block 1
1B	1A	2B	2A	Blk 2
1B	1A	2A	2B	Blk 3
2A	2B	1B	1A	Blk 4

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)

i = a particular main plot treatment

j = a particular block

k = a particular subplot treatment

Steps of Analysis:

Step 1: Construct two way tables of totals.

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

1.2. Block by **Factor A by factor B** total two-way table and calculate factor B totals **Feeds (B)**

Step 2: Compute the correction factor (CF) and sum of squares (SS) for the main-plot analysis

Step 3: Compute the sum of squares for the sub-plot analysis.

Step 4: Compute the F-value for each effect that needs to be tested.

Step 5: Obtain the corresponding tabulated F-value and compare it with the calculated F-value at prescribed level of significance.

Step 6. Compute the two coefficients of variation, one corresponding to the main-plot analysis and another to the sub-plot analysis.

Step 7. Compute mean comparisons

CHAPTER 3

FACTORIAL EXPERIMENTS

3.1. Concepts of Factorial Experiments

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 (**N**) & Phosphorus (**P**) rates:

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

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

Levels of protein supplement (25%, 50%, 75%) (Like; noug cake, groundnut cake) .

Factor and levels:

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 x 3 factorial experiment** with 5 levels of **variety factor (A)** and 3 levels of **spacing factor (B)**. An experiment involving 3 factors (variety, N-rate, weeding method) each at 2 levels is referred as **2 x 2 x 2** or **2³ factors**; 3 refers to the number of factor and 2 refers to level. We have **8 treatment combinations** variety (x, y), N-rate (0, 50 kg/ha), weeding (with weeding or without weeding). The **2³ x 3** is 3 factors each at 2-level and the 4th factor at 3 levels.

If the above **2³** factorial experiment is done in **RCBD**, the correct description of the experiment will be **2³ factorial experiment in RCBD**.

Interaction between factors:

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. 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, Main Effects and Interaction:

Simple effects:

- Simple effect of variety at N_0 : $2-1 = 1$
- Simple effect of variety at N_1 : $4-1 = 3$
- Simple effect of N on variety X: $1-1 = 0$
- Simple effect of N on variety Y: $4-2 = 2$

Main effects:

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)
 $= \frac{1}{2} [(a_0b_1 - a_0b_0) + (a_1b_1 - a_0b_1)] = \frac{1}{2} [(1-1) + (4-2)] = 1$

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**.

$$= \frac{1}{2} (\text{Simple effect of A at } b_1 - \text{simple effect of A at } b_0)$$

$$= \frac{1}{2} [(a_1b_1 - a_0b_1) - (a_1b_0 - a_0b_0)] = \frac{1}{2} [(4-1) - (2-1)] = 1$$

Or

$$= \frac{1}{2} (\text{Simple effect of B at } a_1 - \text{simple effect of B at } a_0)$$

$$= \frac{1}{2} (a_1b_1 - a_1b_0) - (a_0b_1 - a_0b_0) = \frac{1}{2} [(4-2) - (1-1)] = 1$$

In factorial experiments, the following points should be considered:

1. An interaction effect between two factors can be measured only if the two factors are tested together in the same experiment.
2. 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.
3. When **interaction is present**, the simple effect of a factor changes as the level of the other factor changes.

Uses of factorial experiments:

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

Disadvantages:

1. 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.
2. Large factorial experiments are difficult to interpret especially when there are interactions.

3.2. Two Factor Factorial in RCBD

Example: An agronomist wanted to study the effect of different rates of phosphorus fertilizer on two varieties of bean plants. He thought that the varieties might respond differently to fertilizer so he decided to use a factorial experiment with 2- factors.

1. Variety at two level

T_1 = short, bushy and T_2 = tall, erect

2. Phosphorus rate at 3 level

P_1 = none, P_2 = 25kg/ha and 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 **RCBD** with four blocks of six plots each.

Field layout, yield of bean in (Q/ha):

Block I

T ₂ P ₂	T ₂ P ₁	T ₁ P ₁	T ₂ P ₃	T ₁ P ₃	T ₁ P ₂
8.3	11.0	11.5	15.7	18.2	17.1

Block II

T ₂ P ₁	T ₂ P ₂	T ₂ P ₃	T ₁ P ₂	T ₁ P ₁	T ₁ P ₃
11.2	10.5	16.7	17.6	13.6	17.6

Block III

T ₁ P ₂	T ₁ P ₁	T ₂ P ₁	T ₁ P ₃	T ₂ P ₃	T ₂ P ₂
17.6	14.3	12.1	18.2	16.6	9.1

Block IV

T ₁ P ₃	T ₂ P ₂	T ₂ P ₃	T ₂ P ₁	T ₁ P ₂	T ₁ P 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 tables for factors, calculate factor A total, Factor B total and grand total

Variety (Factor A)	Phosphorus (Factor B)			Factor A total
	P1	P2	P3	
T1 (indeterminate or short)	53.9	70.4	72.9	197.2
T2 (determinate or tall)	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	Sum
Total	81.8	87.2	87.9	94.4	351.3

2. Compute C.F, total SS, block SS, treatment SS and Error SS

$$- \text{C.F.} = \frac{G^2}{rab} = \frac{(351.3)^2}{4 \times 2 \times 3} = 5142.15,$$

Where, r is the number of replication (or number of block), a is level of factor A and b is level of factor B

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

- Block SS = $\frac{(81.8)^2 + (87.2)^2 + \dots + (94.4)^2}{6(ab)} - 5142.15 = 13.32$
- Treatment SS = $\frac{(53.9)^2 + (46.9)^2 + \dots + (66.5)^2}{4(r)} - 5142.15 = 221.38$
- Error SS = Total SS – Block SS – 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.

- Factor A (variety) SS = $\frac{\sum A}{rb} - C.F = \frac{(197.2)^2 + (154.1)^2}{4 \times 3} - 5142.15 = 77.40$
- Factor B (P-rate) SS = $\frac{\sum B}{ra} - C.F = \frac{(100.8)^2 + (111.1)^2 + (139.4)^2}{4 \times 2} - 5142.15 = 99.87$
- (A x B) SS = Trt SS – Factor A SS – Factor B SS = 221.38 – 77.40 – 99.87 = 44.11

ANOVA Table

Source	DF	SS	MS	F-cal.	F Table	
					5%	1%
Block	r-1(4-1) = 3	13.32	4.44	7.65**	3.29	5.42
Variety	a-1 (2-1) = 1	77.40	77.40	33.4**	4.54	8.68
Phosphorus (P)	b-1 (3-1) = 2	99.87	49.93	86.09**	3.68	6.36
V X 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				

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

Interpretation of a factorial experiment:

The interpretation of the results of factorial experiment depends on the outcome of the significance tests. If factor A x factor B interaction is significant, the main effects have no real meaning whether significant or not. In our case, since A x B interaction is highly significant, the results of experiment are best summarized in a two – way table means of various A x 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 2 factor factorial experiment.

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

Variety X Phosphate means:

Variety (A)	Phosphorus (B)			Variety mean (A total/rb)
	P1	P2	P3	
T1 (indeterminate or short)	13.47	17.60	18.22	16.43
T2 (determinate or tall)	11.72	10.17	16.62	12.84
Phosphorus mean (B total/ra)	12.59	13.88	17.42	

Standard error of mean differences (Sd) or (SE):

- \bar{Sd} to compare factor A means: $\bar{Sd} A = \sqrt{\frac{2xMSE}{rb}} = \sqrt{\frac{2x0.58}{4x3}} = 0.31Q$
- \bar{Sd} to compare factor B means: $\bar{Sd} B = \sqrt{\frac{2xMSE}{ra}} = \sqrt{\frac{2x0.58}{4x2}} = 0.38 Q$
- \bar{Sd} to compare factor combination (treatment) means: $\bar{Sd} AB = \sqrt{\frac{2xMSE}{r}}$
 $= \sqrt{\frac{2x0.58}{4}} = 0.54Q$

Analysis of Two-Factor CRD without Replication

Assume no interaction between factors

1. Arrange the data:Factor A (Incubation material): **M1, M2, M3**Factor B (Incubation period): **P1, P2, P3, P4**

	M1	M2	M3	Total	Mean
P1	x ₁₁	x ₁₂	x ₁₃	x _{1.}	x _{1./3}
P2	x ₂₁	x ₂₂	x ₂₃	x _{2.}	x _{2./3}
P3	x ₃₁	x ₃₂	x ₃₃	x _{3.}	x _{3./3}
P4	x ₄₁	x ₄₂	x ₄₃	x _{4.}	x _{4./3}
Total	x _{.1}	x _{.2}	x _{.3}	x _{.4}	
Mean	x _{.1/4}	x _{.2/4}	x _{.3/4}	x _{.4/4}	

- $C.F = \frac{G^2}{n(12)}$
- Total sum of square $\sum y_{ij}^2 - C.F$
 $= [(x_{11})^2 + (x_{12})^2 + \dots + (x_{43})^2] - C.F.$
- Sum of squares for M = $\frac{(x_{.1})^2 + (x_{.2})^2 + (x_{.3})^2}{4} - C.F. = \frac{\sum T_i^2}{r} - C.F.$
- Sum of squares for P = $\frac{(x_{1.})^2 + (x_{2.})^2 + (x_{3.})^2 + (x_{4.})^2}{3} - C.F.$
- Interaction (error sum of squares) = Total SS - SS for M - SS for P

ANOVA Table:

Source of Variation	DF	SS	MS	Computed F
M	3-1 = 2	SSM	SSM/2 (m1)	m1/m3
P	4-1 = 3	SSP	SSP/3 (m2)	m2/m3
M x P (error)	2 x 3 = 6	SSE	SSE/6 (m3)	
Total	12-11	TSS		

- $SS_{\bar{x} P} = \sqrt{m3/3}$; $SS_{\bar{d} P} = \sqrt{2m3/3}$
- $SS_{\bar{x} M} = \sqrt{m3/4}$; $SS_{\bar{d} M} = \sqrt{2m3/4}$
- $LSD \alpha = t_{\alpha/2} (6 \text{ d.f.}) s_{\bar{d}}$

CHAPTER 4

COMPARISON BETWEEN 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 (LSD) 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.

1. Least Significant Difference (LSD) Test

It is 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 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, use Fisher's protected LSD test which separates the treatment means only if the ANOVA for treatments shows significance effect 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 treatments means to be compared
3. Compute the LSD value at a level of significant $LSD_{\alpha} = t_{\alpha/2} (n) \times s d$; **where** $s d$ = 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 six 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 ns).

For example:

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 $t(t-1)/2$ 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. Note that we have to separate the treatment means either at 5% or 1%, but not both in the same table. In field experiments, we usually separate at 5% level of significance.

Presentation of data using LSD

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

2. 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 (sd). *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 $= q_{(n, v)} \times \sqrt{(MSE/r)}$

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
<i>Mean stem rust incidence (%)</i> :	50.3	69.0	24.0	94.0	75.0	95.3

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

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

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.

3. Duncan's Multiple Range Test (DMRT)

When we run Analysis of Variance (ANOVA), the results will tell us if there is a difference in means. However, it won't pinpoint which means are different. Duncan's Multiple Range test (DMRT) is a post hoc (second) test to measure specific differences between pairs of means. 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 than the LSD 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). For example, while the LSD might say a difference of means of 6 is significant, the DMRT value might be double that. Technology is usually used to find values for DMRT. By hand, the procedure is essentially the same as that for Fisher's LSD. The main difference is that instead of looking up the critical value in a t-table, we would look in a q-table.

Duncan's Multiple Range Test Example

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

- Note that this test assumes you have run ANOVA and have a significant result.
- Calculate DMRT for the following output:

Variety	A	B	C	D
	17.6	12.0	10.0	8.0
	10.4	9.2	12.0	4.8
	12.0	16.0	9.2	9.6
	8.0	12.0	12.0	4.4
Total	48	49.2	43.2	26.8
Variety mean	12.0	12.3	10.8	6.7
Grand total	167.2			
	SV	Df	SS	MS
	Trts	3	80.2	26.7
	Row	3	46.9	15.6
	Clmn	3	48.8	16.3
	Err	6	2.7	0.45
	Total	15	178.6	

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.

B (12.3) A (12) C (10.8) D (6.7)

The next few steps are to compare the highest mean (12.3) with the lowest (6.7) mean, which have a difference of $12.3 - 6.7 = 5.6$.

Step 2: Calculate σ_d (the standard error of the treatment mean difference) as:

$$\sigma_d = \sqrt{2MSE/r} = \sqrt{2 \times 0.45/4} = \sqrt{0.9/4} = \sqrt{.225} = 0.47$$

Step 3: Look up the q value in this table. With 4 treatments, and 6 degrees of freedom for the error term, the Q value is 4.9.

Step 4: Multiply σ_d (Step 2) by the q-value (Step 3): $0.47 \times 4.9 = 2.3$.

Simply it mean that, $DMRT = q_{0.05}(4, 6) \times \sqrt{2EMS/r} = 4.9 \times 0.47 = 2.3$.

The difference between the highest and lowest means is greater than 2.3, so the highest mean is significantly different from that of the lowest mean.

Step 5: The next few steps are to compare the second highest mean (12) with the lowest mean (6.7), which have a difference of $12 - 6.7 = 5.3$.

Step 6: Look up the q value in this table. With 3 treatments (by excluding the highest mean now, so the q-value will change to 3), and 6 degrees of freedom for the error term, the Q value is 4.34.

Step 7: calculate DMRT VALU as: $DMRT = q_{0.05}(3, 6) \times \sqrt{2EMS/r} = 4.34 \times 0.47 = 2.0$

The difference between the second highest and lowest means is greater than 2.0, so the second highest mean is significantly different from that of the lowest mean.

If we have more values in our table, continue down until we find a non-significant result, if so we can stop at that point.

In general test the difference between treatment means in the following order.

- Largest – Smallest = $12.3 - 6.7 = 5.6$. Compare with SSD (Shortest Significant Difference) or DMRT value at $(t = 4) = 2.2$;
= $d(5.6) > SSD$ at $t = 4(2.2)$; thus the difference is significant at 5% level of significance.
- Largest – 2nd smallest = $12.3 - 10.8 = 1.5$. Compare with SSD at $(t=3) = 2.0$;
= $d(1.5) < SSD$ at $t = 3(2.0)$; thus, the difference is non-significant at 5% level of significance.
- Largest – 2nd largest = $12.3 - 12.0 = 0.3$. compare with SSD at $(t=2) = 1.6$;
= $d(0.3) < SSD$ at $t = 2(1.6)$; thus, the difference is non-significant at 5% level of significance.
- 2nd largest – smallest = $12.0 - 6.7 = 5.3$. Compared with SSD at $(t=3) = 2.0$;
= $d(5.3) > SSD$ at $t = 3(2.0)$; thus, the difference is significant at 5% level of significance.
- 2nd smallest – smallest = $10.8 - 6.7 = 4.1$ compared with SSD at $(t = 2) 1.6$;
= $d(4.1) > SSD$ at $t = 2(1.6)$; thus, the difference is significant at 5% level of significance.
- Etc.

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

SSD (t = 4) = 2.2; SSD (t = 3) = 2.0; SSD (t = 2) = 1.6

Treatments B & D, A & D, C & D are significantly different at 5%, while treatments B & C, B & A, and A & C are not significantly different at 5% level of significance.

Step 8: Present the test result in one of the following two ways.

- Use a line notation if the sequence of results can be arranged according to their ranks.
- 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 5% 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

Mean yields of oat 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 5% 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.

CHAPTER 5

MISSING (PROBLEM) DATA

5.1. 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. An estimate of a missing value does not supply additional information to the experimenter, but it only facilitates the analysis of the remaining data.

Concepts in missing data:

- ☐ Data may be missing because of different reasons, such as;
 - When an animal becomes sick or dies but not due to treatment,
 - When rodents destroyed a plot in field,
 - When a flask breaks in laboratory,
 - When there is an obvious recording error,
- ☐ When missing data in **CRD** experimental design is, it is better to omit the missing treatment rather than trying to estimate it.
 - Because there may not be significant difference on analysis due to the missing observation
- ☐ When missing data in **RCBD** experimental design it is better to look different events;
 - If all of the missing values are on the same block or treatments the simplest solution is to act as if the block or treatment had not been included in the experiment
- ✓ Hence, analyze by using N-1 Block
- ☐ When a single value is missing in **RCBD**, calculate an estimate of the missing value by:

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

Where:

Y= estimate of the missing data

t = # of treatments

r = # of replications

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

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

Go = Grand total of all observed values.

- ☐ The estimated value is entered in the table with the observed values and the **ANOVA** is performed as usual with *one d.f* being subtracted from *both total and error d.f*;

- because the estimated value make no contribution to the error sum of squares

EXAMPLE:

- ❑ In the table given below are dry matter yields (**kg**) of **4-varieties** of forages (Alfalfa, Rhodes, sesbania, vetch) in 4-replications planted in **RCBD** for which one plot yield is missing
- ❑ Estimate the missing value and analyze the data
- ❑ Plot size = 10m x 10 m = 100m²

No	Treatments	I	II	III	IV	Treatment total (Ti)	Treatment mean
1	Alfalfa	18.5	15.7	16.2	14.1	64.5	16.1
2	Rhodes	11.7	Y	12.9	14.4	39(To)	12.9
3	Sesbania	15.4	16.6	15.5	20.3	67.8	16.7
4	Vetch	16.5	18.6	12.7	15.7	63.5	15.9
	Block total	62.1	50.9 (Bo)	57.3	64.5	234.8	

Solution:

A. Estimate the missing value

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

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

B. Enter the estimated value and carry out the analysis

- 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. $CF = \frac{(248.7)^2}{r \times t(4 \times 4)} = 3865.73$
2. Total SS = $\sum_{i=1}^4 \sum_{j=1}^4 Y_{ij} - C.F. = (18.5)^2 + (11.7)^2 + \dots + (13.9)^2 + \dots + (15.7)^2 - C.F. = 79.18$
3. Trt 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. Blocks 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; SS = 79.18 - 31.21 - 9.02 = 38.95

D. Compute the correction factor for bias (B) for treatment sum of square (SS_{TRT})

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

Where, B_o = Total of observed values in blocks (replication) containing the missing data and
 Y = estimated value;

$$B = \frac{[50.9 - (4-1)13.9]^2}{4(4-1)} = \frac{[50.9 - 41.7]^2}{12} = 7.05$$

E. Subtract the computed B value from total SS & treatment SS

- Adjusted treatment SS = treatment SS – B
 $= 31.21 - 7.05 = 24.16$
- Adjusted total SS = Total SS – B
 $= 79.18 - 7.05 = 72.13$

F. Subtract 1 from error d.f and total d.f. and complete the ANOVA table

Source of variation (SV)	Degree of freedom (d.f)	Sum of square (SS)	Mean of square (MS)	Computed F	Tabulated F	
					5%	1%
Treatment	3	24.16	8.05	1.64		
Block	3	9.02	3.01	0.61		
Error	$(t-1)(r-1)-1 = 8$	38.95	4.9			
Total	$rt-1-1 = 14$	72.13				

$$CV = \frac{\sqrt{MSE}}{\text{Grandmean}} \times 100 = \frac{\sqrt{4.9}}{15.65(234.8/15)} \times 100 = 14.3\%$$

When missing data is occurred in **LSD** experimental design, the solution may be as follows
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:

$$\text{Bias (B)} = \frac{[G_o - R_o - C_o - (t-1)T_o]^2}{[(t-1)(t-2)]^2}$$

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 oat 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

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

$$Y = \frac{5(41 + 32 + 37) - (2 \times 205)}{(5-1)(5-2)} = \underline{\underline{11.67}}$$

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

$$\text{CF} = G^2 / t = 217.5^2 / 5 = 43306.25 / 25 = \underline{\underline{1892.25}}$$

$$\begin{aligned} \text{SS}_{\text{TOT}} &= (12^2 + 7^2 + \dots + 9^2) - \text{CF} \\ &= 2072.25 - 1892.25 = \underline{\underline{180}} \end{aligned}$$

$$\begin{aligned} \text{SS}_{\text{ROW}} &= (52.5)^2 / 5 + (40.0)^2 / 5 + \dots + (42)^2 / 5 - \text{CF} \\ &= 1923.85 - 1892.25 = \underline{\underline{31.6}} \end{aligned}$$

$$\begin{aligned} \text{SS}_{\text{COL}} &= (40.0)^2 / 5 + (43.5)^2 / 5 + \dots + (43)^2 / 5 - \text{CF} \\ &= 1898.85 - 1892.25 = \underline{\underline{6.6}} \end{aligned}$$

$$\begin{aligned} \text{SS}_{\text{TRT}} &= (41)^2 / 5 + (33)^2 / 5 + \dots + (61)^2 / 5 - \text{CF} \\ &= 1999.85 - 1892.25 = \underline{\underline{107.6}} \end{aligned}$$

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

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

$$\text{Bias (B)} = \frac{\left[\frac{Go - Ro - Co - (t-1)T_0}{2} \right]^2}{[(t-1)(t-2)]} = \frac{\left[\frac{206 - 41 - 32 - (5-1)37}{2} \right]^2}{[(5-1)(5-2)]} = 1.56$$

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

- Adjusted Treatment SS = Treatment SS – B = 107.6-1.56 = 106.04
- Adjusted Total SS = 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 of variation	DF	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 = $\sqrt{(\text{MSE}) / \text{Grand mean} \times 100}$;

CV = $\sqrt{(3.11)/(206/24) \times 100} = \sqrt{(3.11)/8.58 \times 100} = \mathbf{20.5\%}$

To compare treatment C and D, $\text{LSD}_{5\%} = t_{0.025(11)} \times \text{sd}$;

Where; $\text{sd} = \sqrt{\text{MSE}[1/t + 2/(t-1)(t-2)]} = \sqrt{3.11[1/5 + 2/(5-1)(5-2)]} = \mathbf{1.23}$

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

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)

$$\text{LSD}_{5\%} = t_{0.025(11)} \times \text{sd}, \text{ where } \text{sd} = \sqrt{\frac{2\text{MSE}}{r}} = \sqrt{\frac{2 \times 3.11}{5}} = 1.115$$

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

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.

CHAPTER 6

CHECKING THE ASSUMPTIONS AND TRANSFORMATION OF DATA

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.

Commonly used data transformation methods are:

A. 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; $\log(9+1)$; $\log(10) = 1$

B. 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 subtracting 0.5

C. Arcsine (angular) 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:

$$P = (\sin P')^2 \quad P = (\sin 90^\circ)^2 = 1 = 100\%, P = (\sin 41.55^\circ)^2 = 0.44 \text{ or } 44\%$$

CHAPTER 7

PRACTICING COURSE APPLICATION

7.4. Correlation and regression analysis of data

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 FOUR TYPES:

- 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

Correlation: is used to study the relationship between two types of measurements (variables) on the same individual. It can be shown by different techniques including scatter diagrams.

Correlation coefficient (r): is strength (degree) of correlated linear relationship between variables and the value of r is lied in between -1 and 1. When the value of r is close to 1 or -1; it shows as there is a strong relationship (linear correlation) in between variables or measurements. However the -ve and +ve signs shows the direction of linear lines (correlation). When r=0 means, there is weak or null correlation or uncorrelated.

The model (formula) used to compute r is;

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}] [\sum y^2 - \frac{(\sum y)^2}{n}]}}$$

Example:

	x	y	x ²	y ²	Xy	Summation Notations
	2	5	4	25	10	$\sum x = 8$
	1	3	1	9	3	$\sum y = 16$
	5	6	25	36	30	$\sum x^2 = 30$
	0	2	0	4	0	$\sum y^2 = 74$
Sum	8	16	30	74	43	$\sum xy = 43$

$$r = \frac{43 - (8 \cdot 16)/4}{\sqrt{(30 - 8^2/4)(74 - 16^2/4)}} = \frac{43 - 128/4}{\sqrt{(30 - 16)(74 - 64)}} = \frac{43 - 32}{\sqrt{(14)(10)}} = \frac{11}{\sqrt{140}} = \frac{11}{11.8} = \underline{\underline{0.93}}$$

Since the result is close to +1 and it implies that, the variables x and y are highly positively correlated (the variables has strong linear correlation). It means that if variable one increase and the other variable is too and vice versa.

Regression: 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. The strength (degree) of linear regression is lied in between 0 and 1.

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.

Linear regression:

The relationship between any two variables (independent and dependent) is linear if the change in y is constant as x changes throughout 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: $y = a + bx + e$

Where;

y = is the dependent variable

a = the intercept of the line on the y-axis (the value of y when x is 0)

x = is the dependent variable

b = linear regression coefficient, is the slope of the line or the amount of change in y for each unit change in x

e = error tem

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.

Therefore, its equation is developed based on the predictor (independent) and response (dependent) variables. For a simple calculation regression coefficient (y) can be given as $y = a + bx$, (because e is mostly common term in all observation) where; $a = y - bx$ (y-intercept), $b = SS_{xy}/SS_{xx}$ (slope of a line (regression coefficient) Y is the mean of y and X is the mean of x).

Example:

	x	y	xy	x²	y²
	0	1	0	0	1
	1	5	5	1	25
	2	3	6	4	9
	3	9	27	9	81
	4	7	28	16	49
Sum	10	25	66	30	165
	X=2	Y=5			

$$1. \quad SS_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 66 - \frac{(10 \times 25)}{5} = 66 - 50 = \mathbf{16}$$

$$2. \quad SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 30 - \frac{(10)^2}{5} = 30 - 20 = \mathbf{10}$$

$$3. \quad b = SS_{xy}/SS_{xx} = 16/10 = \mathbf{1.6}$$

$$4. \quad a = Y - bX = 25/5 - (1.6)10/5 = 5 - 3.2 = \mathbf{1.8}$$

Hence, the regression equation is written as $Y = a + bX$; $y = 1.8 + 1.6x$. From this computation we can understand that, for each unit change in x, b measures the rate of change in y (i.e. for each unit change in x the rate of change of y is 1.6 times).