BONGA UNIVERSITY

COLLEGE OF AGRICULTURE AND NATURAL RESOURCES

DEPARTMENT OF ANIMAL SCIENCES

BIOMETRY FOR ANIMAL SCIENCES

Prepared By: Alebel Mulia (Animal Breeding and Genetics)

January, 2020

Bonga, Ethiopia

CHAPTER ONE: INTRODUCTION TO STATISTICS

At the end of the chapter student will able to:

- Explain the concept statistics and biometry
- Explain the concept of experimental design
- Define terminologies of biometry
- Describe types of measure of dispersion

Introduction:

Statistics is a branch of scientific method used to study the planning and designing of data collection, data organization, data analysis, data presentation, data interpretation; and drawing of conclusion based on data. Statistics is the sciences of creating, developing, and applying techniques such that decisions can be made and evaluated in the face of uncertainty. It involves the collection, analysis and interpretation of numerical data. **Biometry** is the application of statistical method to biological and agricultural research problems. It involves the collection of data by means of experiments or surveys conducted according to various principles, and the drawing of conclusion (inferences) from data through the uses of various procedures known as statistical analysis. The field of statistics provides some of the most fundamental tools and techniques of the scientific method:

- Forming hypotheses
- Observational studies and designing experiments
- Gathering data
- Summarizing data
- Drawing inferences from data (Testing hypotheses)

The field of statistics can be divided into:

- **Mathematical Statistics:** the study and development of statistical theory and methods in the abstract; and
- **Applied Statistics:** the application of statistical methods to solve real problems involving randomly generated data, and the development of new statistical methodology motivated by real problems.

- Biostatistics is the branch of applied statistics directed toward applications in t he health sciences and biology. Biostatistics is sometimes distinguished from the field of biometry based upon whether applications are in the health sciences or in broader biology
- Biometry is a branch of statistical science in which statistical methods are applied to biological processes.

Statistics may be classified as follows:

1. Descriptive statistics: these statistics describe the properties of a sample with respect to the given variables. This includes mean, median, mode, percentiles, standard deviation, variance, coefficient of variation, correlation coefficient, etc.

2. Inferential statistics: This includes like standard errors, which are not restricted within the limits of a sample unlike the descriptive statistics, rather go beyond the sample and help to make inferences and generalize them from the sample to the entire population.

3. Prediction statistics: This includes statistics such as regression coefficients, used to predicting the most likely scores in one or more dependent variable from the actual score(s) in one or more independent variables for an individual.

❖ Parameter and Statistic

- Any numerical or descriptive measure of a population is referred to as a parameter. Since a parameter is a measure from the entire population, it is a fixed value. It is a summary value or numerical index like mean, standard deviation, median or variance of a variable for the entire inhabitants.
- Any numerical or descriptive measure of a sample is referred to as a statistic. Usually, a given population parameter is corresponds to a sample statistic. Statistic is a summary value or numerical index like mean, median, standard deviation or variance of scores of a variable in a sample. It is a subset of the entire population.
- ❖ Example: Identify the sample (statistic) and population (parameter)

1. A researcher wants to estimate the average height of women aged 20 years or older. From a simple random sample of 45 women, the researcher obtains a sample mean height of 63.9 inches.
   - The parameter is the average height of all women aged 20 years or older, while
   - The statistic is the average height of 63.9 inches from the sample of 45 women.

2. A nutritionist wants to estimate the mean amount of sodium consumed by children under the age of 10. From a random sample of 75 children under the age of 10, the nutritionist obtains a sample mean of 2993 milligrams of sodium consumed.
   - The parameter is the mean amount of sodium consumed by children under the age of ten, while
   - The statistic is the mean of 2993 milligrams of sodium obtained from the sample of 75 children.

The different measurement of variability in Parameter and Statistic

Measure	Statistic	Parameter
Mean	$\bar{X}$ |	μ
Standard deviation	s, s_x	σ , σ_x
Variance	s^2	σ^2
Standard error of the mean	$s_{\bar{X}}$	$\sigma_{\bar{X}}$
Correlation coefficient	γ	ρ

Measures of dispersion or variability

A measure of dispersion is an expression of variability in a data set or how a value is spread around the center. A measure of dispersion has to be reported with average values because it alone does not adequately describe a data set. *Dispersion measures:* range, mean absolute deviation, variance, standard deviation and coefficient of variation.

Range: is the difference between the highest and lowest value in a data set. However, when expressed as a difference, mean value should be reported together with it, otherwise the reader will not understand where the actual values lie. Range over emphasizes the true dispersion in the presence of outlier values because it only takes into consideration two values in the data set. This idea bought into attention other measures of dispersions discussed below.

Mean absolute deviation: is an expression of dispersion as a deviation of each value from the mean. Since the sum of all deviation from the mean is equal to zero, to avoid the sign effect absolute value of the deviations are taken and the formula becomes:

Mean Absolute Deviation $= \dfrac{\sum / xi - \overline{X} /}{n}$ **or**

$$\frac{\text{Total absolute deviation}}{\text{sample size}}$$

This measure of dispersion, although adequate to describe dispersion in a data set, its theoretical distribution is poorly known and hence it has no use in higher statistics so need other.

Variance and standard deviation

Variance: When expressing dispersion as a deviation from the mean, another method of eliminating the sign effect is to square the deviations. Then dividing the sum of the squared deviations by the total number of data points gives a measure of dispersion known as **Variance**.

Population variance $\sigma^2 = \sum (xi - \mu)^2/N$

Sample variance $S^2 = \sum \left(xi - \bar{X}\right)^2/n-1$

Variance is also termed as **Mean square** (MS) since it is the mean of the total squared deviation of each value from the mean. i.e.,

When abbreviated, the expression can be written as **S^2 = MS = TSS/DF**

Sample variance (S^2) is an estimator of the population variance (σ^2). The denominator for the sample variance is n-1, referred to as the degrees of freedom rather than n (sample size) because if the denominator is sample size, S^2 underestimates the true dispersion. Variance expresses the dispersion in the data values in squared unit. Since variance expresses deviation in squared unit, it is not easily graspable by **common** users.

Therefore the positive square root of variance is used to express deviation in linear units and this gives a quantity called standard deviation (S)

- Sample standard deviation = $S = \sqrt{S^2} = \sqrt{\sum \left(xi - \bar{X}\right)^2/n-1}$

For the 1996 class students score, S = √214.29 = 14.64%. i.e., in the 1996 class, the score of a given student deviated from the class average by 14.64%. Table 2.1 this is much easier to understand since it is expressed in similar units as the values in the data set.

For data recorded in frequency table, $S^2 = [\sum f_i \left(xi - \bar{X}\right)^2]/n-1$

Variance is the most preferred measure of dispersion for it satisfies two important qualities. Firstly it takes all values in the data set in its computation and therefore it is not biased as range.

Secondly its probability distribution is well understood and therefore this makes it indispensable in higher statistics.

The Coefficient of variation (CV)

The Coefficient of variation is defined as CV = $\frac{s}{x}$ x100. It describes variability in the data set as a proportion to the mean. CV has no unit and it is a relative measure of variability. It can be used to compare the relative dispersion (variability) of different data sets measured in different units as well as magnitudes. To know the dispersion of data set to the mean value using of CV is better than using variances of the two data sets.

Table 1.2 Biostatistics final score of 1996 and 1997 class students (x/100)

No of student	**Score (X/100)**	
	1996 class	1997 class
1	95	85
2	90	82
3	85	80
4	80	78
5	70	75
6	65	70
7	60	68
8	55	65
Total (∑ X)	**600**	**603**
Mean	**75**	**75.4**
Range	**55-95**	**65-85**
Variance (s^2)	**214.29**	**50.84**
Standard deviation (S)	**14.64**	**7.13**

For example consider the final score on biostatistics (%) of two classes of students, 1996 and 1997 class displayed by Table 1.2. The mean score for the 1996 and 1997 class is 75.0% and 75.4 %, respectively. Thus reporting the mean value alone misleads that the two class students are comparable in performance. On the other hand, if some measure of dispersion like Range is reported together with the mean value, the reader would understand that the two class students are widely different in performance. i.e., the 1996 class students are highly varied in performance but the 1997 class students are relatively close to each other. However, the reader cannot get this idea by inspecting the mean value alone.

Definition of some basic Terminologies

Population: It is a totality of all subjects or group of elements possess certain common characteristics that are being studied. Sample: is a subgroup or a part of a population selected by some methods in order to estimate the characteristic of a population.

Data: Data are the material with which statisticians work. They are records of measurement, counts or observations. Examples of data are records: weights of calves, milk yield in lactation of a group of cows, sex (Male or female), and blue or green color of eyes.

Data may be classified into primary data and secondary data depending on the source of data.

1. Primary data: are data which are collected from the units or individuals respondents directly for the purpose of certain study or investigation are known as primary data.
2. Secondary data: are data, which had been collected by certain people or agency, and statistically treated.

Based on the nature of the characteristic observed, the statistical **data** are classified as attribute data and measurement data.

- What is attribute and measurement data?

Attribute data: The variables that cannot be measured but expressed qualitatively, *e.g.* coat color type of sheep: white, black, red; sex of animals could be male or female, blood group of humans could be A, AB, O, etc.

Measurement data: the Variables that can be expressed in a numerical order. E.g. height, grain yield, score of students, number of seeds per pod, number of plants in a quadrant, number of students taking the course biometry, etc.

Variable: A set of observations on a particular character is termed as a variable and data are the values of a variable. For example, variables denoting the data are weight, age, milk yield, sex, and eye color. **Variables** can be defined as quantitative (numerical) and qualitative (attributive, categorical, or classification).

I. **Quantitative variables:** are variables that assume values of the *measurable quantity* or the *amount* of something. Quantitative variables have values expressed as numbers and the differences between values have numerical meaning. Examples of quantitative variables are weight of animals, litter size, temperature or time. They also can include ratios of two numerical variables, count data, and proportions.

A quantitative variable can be continuous or discrete.

A continuous variable can take on an infinite number of values over a given interval. Its values are real numbers. Examples milk yield or weight, body weight, Body length horn length, **a discrete** variable is a variable that has countable values, and the number of those values can either be finite or infinite. Its values are natural numbers or integers. Examples litter size, number of laid eggs per month, number of children in a family, the number of students in a class room.

II. **Qualitative variable:** are the ones that assume values that are **not numerical** but can be categorized. Qualitative variables have values expressed in categories. Data on gender, religious affiliation, type of occupation, type of land, area of residence, eye color or whether or not an animal is ill etc., are categorical data. A qualitative variable can be an ordinal or nominal. An ordinal variable has categories that can be ranked. Example of an ordinal variable is calving ease scoring. A nominal variable has categories that cannot be ranked. Examples of nominal variables are identification number, color or gender

Scale of measurement: The type of classification, how variables are categorized, counted and measured-uses measurement scales and four common types of scales are used:

1. **Interval Variables:** are continuous measurements that may be either positive or negative and follow a linear scale.
2. **Ordinal Variables:** measurements that may be ordered according to magnitude.
3. **Nominal Variables:** are those in which the number represents the state of the variable.
4. **Ratio Variables:** are positive measurements in which the distinction between two numbers is constant if their ratio is constant.

Nominal level (Is A different with B?)	Interval level data (By how many units do A and B differ?)	Ordinal level data (Is A bigger than B?)	Ratio level data (How many times B is bigger than A?)
Marital status	Temperature	Grade (A,B,C,D,F)	Height
Gender (male, female)		Rating scale (poor, good, excellent)	Weight
Eye color (blue, brown, green, hazel)		Judging (First place, 2nd place, etc.)	Time
Political affiliation		Stage of disease	Age
Religious affiliation		Severity of disease	Distance

Definition and meaning of Design of Experiments

The planning of comparative experiments and the consequent collection and analysis of observation may be separated into three phases:

1. The choice of the treatment to be compared, observations to be made and experimental units to be used.
2. The method of assigning treatments to experimental units and the decision as how many units are to be used.
3. The tabulation, statistical analysis and interpretation of data.

- **Design**: a plan to collect observation.
- **Research design:** The research hypothesis, treatment design, and the experiment or observational study design constitute the research design.
- *Experimental design* is the process of planning a study to meet specified objectives.
- **An experiment** can be defined as planning of 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 as an experiment.
- **Experimental material**: the "stuff" we are applying treatments to (field plots, cows, varieties, etc.,)
- **Experimental unit**: the smallest unit or division of the experimental material such that any two units are likely to receive different treatments.
- **Treatment**: a set of conditions brought about under the control of the experimenters in order to evaluate the effects of these conditions on some experimental material.
- **Sampling unit**: A sub division of the experimental unit. Any 2 sampling units are always subject to the same treatment.
- **Control treatment**: is a necessary bench mark treatment to evaluate the effectiveness of experimental treatments.
- **Experimental error**: observed variation among experimental units which receive the same treatment.
- **Sampling error**: observed variation among sampling units within an experimental unit.

Classification of Agricultural Experiments

Animal research comprises both experiments and surveys conducted either on-station, or on-farm or in the field. Such research has to be properly designed, conducted, statistically analyzed and interpreted.

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

1. Introduction to the problem,
2. Statement of the hypotheses,
3. Description of the experimental design,
4. Collection of data (running the experiment),
5. Analysis of the data resulting from the experiment, and
6. Interpretation of the results relative to the hypotheses

Principles of Agricultural Experiments

(Replication, randomization and Local control or Blocking)

Planning and designing experiment to obtain appropriate data and drawing inference out of the data with respect to any problem under investigation is known as ***design and analysis of experiments***. This might range anywhere from the formulation of the objectives of the experiment in clear terms to the final stage of the drafting reports incorporating the important findings of the enquiry. Before attempting to analyze data, it is necessary to understand how the experiment is designed and from which the data have been obtained, If the experiment has not been designed well then it may not be possible to undertake a satisfactory analysis of the data.

Before designing an experiment it is important that the objectives for the experiment are clear, well-defined, realistic and relevant.

I. **Clear**. If the objectives are vague, it will be difficult to know how to go about planning an experiment.
II. **Well-defined**. If the objectives are not carefully stated then it will not be clear what hypotheses are to be evaluated.
III. **Realistic**. The researcher needs to be confident that an experiment can be designed that meets the objectives.
IV. **Relevant**. The objectives for the experiment need to be relevant to the problem in hand. In other words the researcher will be a step nearer to solving the problem once he/she has the results from the experiment

The purpose of designing an experiment is to increase the precision of the experiment. In order to increase the precision, we try to reduce the experimental error. For reducing the experimental error, we adopt certain techniques. These techniques form the basic principles of experimental designs. The basic principles of experimental designs are:

I. Replication,
II. Randomization and
III. Blocking (local control).

Replication: The repeated application of the treatments under investigation is known as replication. If the treatment is applies only once we have no means to know the variation in the result of a treatment. Only when we repeat the application of the treatment several times we can estimate the experimental error.

Randomization: All treatments have equal chances of being allocated to different experimental unit is known as ***randomization***. The actual procedure of randomization will vary according to the design we adopt.

Blocking or Local control: Putting experimental units on the plots which is similar in the same group (generally referred as block) and by assigning all treatments into each block separately and independently, variation among blocks can be measured and removed from experimental error. In field experiments, substantial variation within an experimental field can be expected; significant reduction in experimental error is usually achieved with the use of proper blocking. Blocking is an important component in almost all experimental designs.

Planning of an experiment properly is very important to:

- ✓ Ensure that the right type of data,
- ✓ Have sufficient sample size and
- ✓ Acquire available power to answer the research questions

During designing of an experiment, the **following steps to be performed by a researcher:**

1. Define the problem and the questions to be addressed.
2. Define the population of interest.
3. Determine the need for sampling.
4. Define the experimental design.

CHAPTER TWO: COMMON DESIGNS OF AGRICULTURAL EXPERIMENTS

At the end of the chapter student will able to:

- Understand the concept of agricultural experiments
- List and explain the common design of agricultural experiments
- Explain source of variation for experimental unit for each design
- Know randomization procedure of treatment on experimental unit
- Compare different treatment mean for all design

Introduction:

An experiment is a series of tests in which the input variables are changed according to a given rule in order to identify the reasons for the changes in the output response. An experiment aims at predicting the outcome by introducing a change of the conditions, which is represented by one or more independent variables, also referred to as "input variables" or "predictor variables." Experimental design is a way to carefully plan an experiment to obtain new facts, or to confirm or refute the results of previous experiments. It is the design of any experiment that aims to describe and explain the variation of experimental unit under treatment that is hypothesized to reflect the variation.

They can be broadly classified as a single-factor experiments and mulit-factor (factorial) experiments. The single-factor experiments can be grouped as complete ***block designs*** and CRD. When the treatments consists different levels of a single variable factor and all other factors are kept at a single prescribed level, it is known as a ***single-factor experiment***. When ANOVA design deals with evaluation of the simultaneous effects of two or more factors on the response variable, it is called **multi-factor ANOVA or factorial**.

To select appropriate experimental design that best suit objectives of an experiment, the following must be known.

- Type and number of treatments involved.
- The degree of precision desired.
- The size of "uncontrollable" variation.

In selecting appropriate design, a researcher must be selecting the simplest design that controls variability and attains the desired precision. There are many types of experimental designs. Commonly used designs are, CRD, RCBD, Latin square, Split plot, and factorial experiment etc.

2.1. Completely Randomized Design (CRD)

2.1.1. Description of CRD

Completely Randomized Designs are used to study the effects of a single factor without considering any other variables. It is a flexible design and one of the least used designs but could be helpful in a situation where experimental area is believed to be uniform and few treatments are examined. The statistical analysis is also simple even when there are unequal replications or missing values. CRD is the basic single-factor experiment that all other designs like RCBD and Latin square designs stem from it. Completely randomized design is only appropriate for experiments with homogeneous experimental units such as laboratory experiments. It is rarely used in field experiments due to large variation among experimental plots.

2.1.2. Advantages and Disadvantages of CRD

Advantages of a CRD

1. Very flexible design (i.e. number of treatments and replicates is only limited by the available number of experimental units).
2. Statistical analysis is simple
3. Loss of information due to missing data is small
4. It is best suited for experiments with a small number of treatments.

Disadvantages

1. If experimental units are not homogeneous, there may be a loss of precision.
2. Usually the least efficient design unless experimental units are homogeneous.
3. Not suited for a large number of treatments.

2.1.3. Layout of CRD

Treatments are assigned to experimental units completely at random. Every experimental unit has the same probability of receiving any treatment. Placement of the treatments on the experimental units along with the arrangement of experimental units is known as the **layout of the experiment**.

The randomization procedure of treatments on experimental units will be as follows:

1. Determine the total number of experimental units.
2. Assign a plot number to each of the experimental units starting from left to right for all rows.
3. Assign the treatments to the experimental units by using random numbers.

Suppose that there are T treatments on the experimental units, namely, T_1, T_2… T_t and are replicated (**r**) times each, we require, txr=n experimental units if **replication is equal and factor is one**. In case of unequal replications, the number of experimental units required will be:

$u_1 + u_2 + ... + u_t = n$ Or the sum of all observation;

Where u_1……u_t stands for experimental units

For example, suppose there are five treatments each with four replications. We need 20 experimental units.

The 20 units are numbered as follows.

$U1_{T4}$	$U2_{T3}$	$U3_{T4}$	$U4_{T1}$	$U5_{T3}$
$U10_{T1}$	$U9_{T1}$	$U8_{T3}$	$U7_{T2}$	$U6_{T3}$
$U11_{T5}$	$U12_{T4}$	$U13_{T2}$	$U14_{T2}$	$U15_{T4}$
$U20_{T5}$	$U19_{T2}$	$U18_{T1}$	$U17_{T5}$	$U16_{T5}$

That is, 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:

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

2.1.4. Analysis and interpretation of CRD

The ANOVA model for CRD is

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

Where *yij*= observation *j* in group or treatment *i*

μ = grand mean

t_i = treatment effect

e_{ij} = random error.

The sources of variation are, therefore, "between treatments" and "within treatments" or "experimental error". The computation procedure is given below.

	Treatments			….		
	1	2	3	…..	T	
	Y_{11}	Y_{21}	Y_{31}	…..	Y_{t1}	
	Y_{12}	Y_{22}	Y_{32}	…..	Y_{t2}	
	Y_{13}	Y_{23}	Y_{33}	…..	Y_{13}	
	-	-	-	-	..	
	-	-	-	-	..	
	-	-	-	-	..	
	-	-	-	-	..	GT=Grand Total
Total	T_1	T_2	T_3		T_t	GT
No. of replication	r_1	R_2	r_3		r_t	N

Then homogeneity of variance is tested. If the variances are homogeneous proceed with the analysis. In case of heterogeneous variances, the data transformation has to be done and then the analysis should be continued. Next, the required sums of squares are computed.

The test of hypothesis

Hypotheses are about the differences between population means. A null hypothesis H_0 and an alternative hypothesis H_1 are stated:

H_0: $\mu_1 = \mu_2 = ... = \mu a$, the population means are equal

H_1: $\mu i \neq \mu i'$, for at least one pair (i,i'), the means are not equal

The hypotheses can also be stated:

H_0: $\tau_1 = \tau_2 = ... = \tau a$, there is no difference among treatments, i.e. there is no effect of treatments

H_1: $\tau i \neq \tau i'$, for at least one pair (i,i') a difference between treatments exists.

The analysis of variance table, and mean squares and F are computed, as given in the table below.

Table of Analysis of variance for CRD with t treatments

ANOVA				
Source	d.f.	SS	MS	F- value
Treatment	t-1	SSt	MSt = SSt/(t-1)	F=MSt/MSE
Error	t(r-1)	SSE	MSE = SSE/ (r(r-1)	
Total	rt-1	SST		

The total variation is partitioned into two components:

a. Variation among treatment means

b. Variation among plots within treatments (error).

If F- calculated value is greater than the F-tabulated value, we reject the null hypothesis and conclude that there are significant differences among the treatment means.

2.1.5. CRD with equal and unequal replication

Equal replications

The steps involved in the analysis of variance for data from a CRD experiment with an equal number of replications are given below.

Step 1. Group data by treatments and calculate treatment totals (T), and grand total (G).

Step 2. Using t (treatments) and r (replications), determine the degrees of freedom (d.f.) for each source of variation as follows:

- ® Total df = (r) (t) -1
- ® Treatment d.f. = t-1
- ® Error d.f. = t(r-1)

❖ The error d.f. can also be obtained through subtraction as:

- ® [Error d.f. = Total d.f. – Treatment d.f.

Step 3. Using Xi to represent the measurement of the i^{th} plot, Ti as the total of the i^{th} treatment, and n as the total number of experimental plots (i.e., n = (r) (t)),

- Calculate the correction factor, and the various sums squares (SS) as:

  - ® Correction factor (C.F.) = $\frac{(GT)^2}{n}$
  - ® Total SS = $\sum_{i=1}^{n} X_i^{\,2} - C.F.$
  - ® Treatment SS = $\frac{\sum_{i=1}^{t} T_i^2}{r} - CF$
  - ® Error SS = Total SS – Treatment SS

Step 4. Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

$$\text{Treatment MS} = \frac{TSS}{t-1}$$

$$\text{Error MS} = \frac{ESS}{t(r-1)}$$

Step 5. Calculate the F value for testing significance of the treatment difference as:

$$F = \frac{TMS}{EMS}$$

Step 6. Obtain the F values from the table, with f_1 = treatment *df* = (t-1) and f_2 = error *df* = t(r-1) by using given significance level (5% or 1%).

Step 7. Construct ANOVA table and enter all the values computed in steps 2 to 6 and Compare the computed F values of step 5 with the tabular F values of step 6, and decide to reject Ho if Fcal ≥ Ftab or accept Ho if Fcal ≤ Ftab.

Source of variation	*Df*	SS	MS	Computed F	Tabular F		Decision
					5%	1%	
Treatment							
Error							
Total							

Step8. Conclusion: Conclude the significance among treatments using the following rules:

- ✓ If F Calculated > F tabulated at 1% ** (Highly significant)
- ✓ If F Calculated > F tabulated at 5% * (Significant)
- ✓ If F Calculated <= F tabulated at 5% ns (Non-significant)

Step 9. Compute the grand mean and the coefficient of variation (CV) as follows:

Grand mean = G/n

$$CV = \frac{\sqrt{EMS}}{Grandmean} x100$$

CV expresses the experimental error as percentage of the mean; thus, the higher the *CV* value, the lower is the reliability of the experiment. It is usually placed below the analysis of variance table.

Example 1: Consider that the effects of insecticides for the control of brown plant hoopers and stem borers on the grain yield of rice which was evaluated by investigator.

Treatment	Grain yield, kg/ha			
Dol-Mix (1kg)	2537	2069	2104	1797
Dol-Mix (2kg)	3366	2591	2211	2544
DDT +γ-BHC	2536	2459	2827	2385
Azodrin	2387	2453	1556	2116
Dimecron-Boom	1997	1679	1649	1859
Dimecron –Knap	1796	1704	1904	1320
Control	1401	1516	1270	1077

Step 1. Group data by treatments and calculate treatment totals (T), and grand total (G).

Treatment	Grain yield, kg/ha				**Treatment Total**
Dol-Mix (1kg)	2537	2069	2104	1797	**8507**
Dol-Mix (2kg)	3366	2591	2211	2544	**10712**
DDT +γ-BHC	2536	2459	2827	2385	**10207**
Azodrin	2387	2453	1556	2116	**8512**
Dimecron-Boom	1997	1679	1649	1859	**7184**
Dimecron –Knap	1796	1704	1904	1320	**6724**
Control	1401	1516	1270	1077	**5264**
Grand T (G)					**57110**

Step 2.Determine the degrees of freedom (d.f.) for each source of variation as follows:

Total df = (r) (t) -1 = (4) (7) -1 = 27

Treatment d.f. = t-1 = 7-1 = 6

Error d.f. = t(r-1) = 7(4-1) = 21

The error d.f. can also be obtained through subtraction as:

Error d.f. = Total d.f. – Treatment d.f. = 27-6 = 21.

Step 3.Calculate the correction factor and the various sums of squares (SS) as:

$$\text{C.F.} = \frac{(GT)^2}{n} = \frac{(57{,}110)^2}{(4)(7)} = 116{,}484{,}004$$

$$\text{Total SS} = \sum_{i=1}^{n} X_i^{\,2} - C.F. = \left[(2537)^2 + (2069)^2 + ... + (1270)^2 + (1077)^2\right] - 116484004 = 7577412$$

$$\text{Treatment SS} = \frac{\sum_{i=1}^{t} T_i^{\,2}}{r} - CF = \frac{(8507)^2 + (10712)^2 + ... + (5264)^2}{4} - 116484004 = 5587174$$

Error SS = Total SS – Treatment SS =7577412-5587174=1990238

Step 4. Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

$$\text{Treatment MS} = \frac{TrSS}{t-1} = \frac{5587174}{6} = 931196$$

$$\text{Error MS} = \frac{ESS}{t(r-1)} = \frac{1990238}{(7)(3)} = 94773$$

Step 5. Calculate the F value for testing significance of the treatment difference as:

$$F = \frac{TMS}{EMS} = \frac{931196}{94773} = 9.83$$

Step 6. Obtain the F values from the table, with f_1 = treatment *df* = (t-1) and f_2 = error *df* = t(r-1). For our example, the tabular F values for 5% level= $F_{0.05}$ (t-1, t(r-1**)** = $F_{0.05}$ (6, 21**)** = 2.57 and for 1% $F_{0.01}$ (6, 21) = 3.81.

Step 7. Construct ANOVA table and enter all the values computed in steps 2 to 6 and Compare the computed F values of step 5 with the tabular F values of step 6, and decide to reject or accept $H_{o.}$

Source of variation	*Df*	SS	MS	Computed F	Tabular F		Decision
					5%	1%	
Treatment	6	5587174	931196	9.83**	2.57	3.91	accept H_1
Error	21	1990238	94773				
Total	27	7577412					

Step8. Conclusion

From this example, the F calculated is greater than F tabulated (** = Significant at 1% level)

So, **the grain yield of rice is highly different (p<0.01) under different insecticide treatments**.

Step 9. Compute the grand mean and the coefficient of variation (*cv*) as follows:

Grand mean = $\frac{G}{n} = \frac{57110}{28} = 2040$

$$CV = \frac{\sqrt{EMS}}{Grandmean} x100 = \frac{\sqrt{94773}}{2040} x100 = 15.1\%$$

The *CV* indicates the degree of precision with which the treatments are compared and is a good index of the variability of the experiment. It expresses the experimental error as percentage of the mean; thus, the higher the *CV* value, the lower is the reliability of the experiment. It is usually placed below the analysis of variance table.

Unequal replications

Because the computational procedure for the CRD is not complicated when the number of replications differs among treatments, the CRD is commonly used for studies where the experimental material makes it difficult to use an equal number of replications for all treatments.

Some examples of these are:

Animal feeding experiment when the number of animals for each breed is not the same.

Experiments where some units lost or damaged

Example 2

A variety trial on green gram was conducted in a CRD with five varieties, V1, V2, V3, V4, V5, and 3, 4, 5, 4 and 4 replications, respectively. The results are presented in Table 3. The net plot size was 5 x 3.5 meters.

Grain yield of green gram, kg/plot.

	Varieties					Grand total	Grand mean
	1	2	3	4	5		
	1.6	2.5	1.3	2.0	1.6		
	1.2	2.2	0.9	1.5	1.0		
	1.5	2.4	0.8	1.6	0.8		
		1.9	1.1	1.4	0.9		
			1.0				
Variety total	4.3	9.0	5.1	6.5	4.3	29.2	1.46

$$CF = \frac{(29.2)^2}{20} = 42.6320$$

Total SS = $(1.6)^2 + (1.2)^2 + ... + (0.9)^2 - CF$

= 47.8400-42.6320 = 5.2080

Variety SS = $\left[\frac{(4.3)^2}{3} + \frac{(9.0)^2}{4} + \frac{(5.1)^2}{5} + \frac{(6.5)^2}{4} + \frac{(4.3)^2}{4} \right] - CF$

= 46.8003-42.6320

= 4.168

Error SS = Total SS – Variety SS

= 5.2080-4.1683=1.0397

Calculate the F value for testing significance of the treatments

ANOVA table for the data presented.

Source of variation	Df	SS	MS	Computed F	Tabular F 5% 1%
Variety	4	4.1683	1.0421	15.037**	3.06 4.89
Error	15	1.0397	0.0693		
Total	19	5.2080			

CV= $cv = \frac{\sqrt{ErrorMS}}{Grandmean} X100 = 18.03$

** = significant at 1% level.

2.2. Randomized Complete Block Design (RCBD)

2.2.1. Description RCBD

RCBD is the most widely used design in agricultural research. The experimental units are divided into homogeneous groups of material (called Blocks) each of which constitutes a single replication of the experiment. The RCBD has equal block sizes which contain all the treatments. A randomized complete block design is used when experimental units can be grouped in blocks according to some defined source of variability before assignment of treatments. Blocks are groups that are used to explain another part of variability, but the test of their difference is usually not of primary interest. The number of experimental units in each block is equal to the number of treatments, and each treatment is randomly assigned to one experimental unit in each block. The precision of the experiment is increased because variation between blocks is removed in the analysis and the possibility of detecting treatment effects is increased.

The characteristics of randomized complete block design are:

1. Experimental units are divided to *a* treatments and *b* blocks.
   Each treatment appears in each block only once.
2. The treatments are assigned to units in each block randomly.

This design is balanced, each experimental unit is grouped according to blocks and treatments, and there is the same number of blocks for each treatment. Data obtained from this design are analyzed with a two-way *ANOVA*, because two ways of grouping, *blocks and treatments*, are defined.

☞ Animals are most often grouped into blocks according to initial weight, body condition, breed, sex, stage of lactation, litter size, etc. It is important that during the experiment all animals within a block receive the same conditions in everything *except treatments*.

2.2.2. Blocking

In many experiments it is recognized in advance that some experimental units will respond similarly, regardless of treatments. For example, neighboring plots will be *more similar than those further apart*, heavier animals will have *different gain* than lighter ones; measurement on the same day will be more similar compared to measurements taken on different days, etc. In these cases experimental designs should be able to account for those known sources of variability

by grouping homogeneous units in blocks. In this way, the *experimental error decreases*, and the possibility of finding a difference between treatments increases.

☞ Consider for example that the aim of an experiment is to compare efficiency of utilization of several feeds for pigs in some region. It is known that several breeds are produced in that area. If it is known that breed does not influence efficiency of feed utilization, then the experiment can be designed in a simple way: randomly choose pigs and feed them with different feeds. However, if an effect of breed exists, variability between pigs will be greater than expected, because of variability between breeds. For a more precise and correct conclusion, it is necessary to determine the breed of each pig. Breeds can then be defined as blocks and pigs within each breed fed different feeds.

2.2.3. Advantages and disadvantage of blocking

Advantages

1. Improves precision (relative to CRD)

Effective blocking reduces S^2_e, thus resulting in greater precision.

2. Flexible

- Any number of treatments replicates
- Any number of block replicates
- Extra replications for certain treatments may be included (2x, 3x…)
- Not all block need to contain the same number of units.

3. The scope of inference is increased and block means provide a comparison of the differences among blocks.

Disadvantages

1. Certain assumptions may be required for some tests of hypothesis

2. Block * treatment interaction may make interpretation of treatment effects more difficult

3. Blocking for a single factor may not provide sufficient error control (precision).

4. The gain in precision due to blocking generally decreases as the number of experimental units in a block increases.

5. Requires some prior knowledge about variability of experimental units for successful blocking.

Where should the RCBD used?

1. Field experiments

2. Experiments where lack of uniformity information is available.

3. When to compare the treatments over a wide range of experimental material to increase generalization.

4. If more than one source of heterogeneity exists in the experimental material, sources may be confounded or multifactor blocking may be required.

2.2.4. Randomization and layout

Suppose the experiment was aimed to determine the effect of three treatments (T1, T2 and T3) on a certain variable and the experimental materials are heterogeneous. So, before the start of the experiment these experimental materials are measured, and ranked according to their measured values, and assigned to four blocks. The three highest measured value materials are assigned to block I, the three next to block II, etc. In each block there are three experimental units to which the treatments are randomly assigned. Therefore, a total of 12 experimental units are used. The identification numbers are assigned to experimental materials in the following manner:

Block	Experimental material number
I	1, 2, 3
II	4, 5 , 6
III	7 , 8 , 9
IV	10,11, 12

In each block the treatments are randomly assigned to steers.

	Block			
	I	II	III	IV
Exp. Mat. No. (Treatment)	No. 1 (T_3)	No. 4 (T_1)	No. 7 (T_3)	No. 10 (T_3)
	No. 2 (T_1)	No. 5 (T_2)	No. 8 (T_1)	No. 11 (T_2)
	No. 3 (T_2)	No. 6 (T_3)	No. 9 (T_2)	No. 12 (T_1)

When an experiment is finished, the data can be rearranged for easier computing as in the following table:

	Block			
Treatment	I	II	III	IV
T_1	y_{11}	y_{12}	y_{13}	y_{14}
T_2	y_{21}	y_{22}	y_{23}	y_{24}
T_3	y_{31}	y_{32}	y_{33}	y_{34}

2.2.5. ANOVA Model for simple RCB Design, and Interpretation of Results

Model: $Y_{ij} = \mu + t_i + b_j + e_{ij}$

Where, y_{ij}= an observation in treatment *i* and block *j*

μ = The overall mean

t_i = The i^{th} treatment effect

b_j = The j^{th} block effect

e_{ij} = The error term

Partitioning Total Variability

- In the randomized complete block design the total sum of squares can be partitioned to block, treatment and residual sums of squares:

$$SSTOT = SSTRT + SSBLK + SSRES$$

The corresponding degrees of freedom are:

$$(ba - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1)$$

☞ Compared to the one-way *ANOVA*, the residual sum of squares in the two-way *ANOVA* is decreased by the block sum of squares. Namely:

$$SS'RES = SSBLK + SSRES$$

Where: *SSRES* = the two-way residual sum of squares (the experimental error for the randomized complete block design)

SS'RES = the one-way residual sum of squares

The consequence of the decreased residual sum of squares is increased precision in determining possible differences among treatments.

The total variance is divided into three source of variation

1. Between block
2. Between treatments
3. Error

The required sum of squares (SS) are obtained as follows:

$$CF = \frac{(G.T.)^2}{r.t}$$

Total SS = $\sum Y^2_{ij} - CF$

Block SS = $\frac{1}{t}\sum B^2_{\ j} - CF$

- Note that the number of observations in a block is equal to the number of treatments.

Treatment SS = $\frac{1}{b}\sum T^2_{\ i} - CF$

- Note that the number of observations in a treatment is equal to the number of blocks.

Error SS = Total SS-Block SS-Treatment SS

With these result the analysis of variance table is completed. The form of the ANOVA table of RCBD with *t* treatments and *b* blocks is given in the following table.

Source	**Df**	**SS**	**MS**	**Computed F**	**Tabular F**		**Decision**
					5%	**1%**	
Blocks	b-1	BSS	BMS = BSS/b-1	BMS/EMS			
Treatment	t-1	TrSS	TrMS = TrSS/t-1	TrMS/EMS			
Error	(b-1)(t-1)	ESS	EMS = ESS/(b-1)(t-1)				
Total	bt-1	Tot SS					

Hypotheses Test - *F* test

☞ The hypotheses of interest are to determine if there are treatment differences. The null hypothesis H_0 and alternative hypothesis H_1 are stated as follows:

*H*0: $\tau 1 = \tau 2 = ... = \tau a$, there are no differences among treatments

*H*1: $\tau i \neq \tau i'$, for at least one pair (*i*,*i*') a difference between treatments exists

→ For an α level of significance H_0 is rejected if F c*alculated* $> F\alpha$, (*a*-1),(*a*-1)(*b*-1), that is, if the calculated *F-value* from the sample is greater than the critical *F-value*.

Compute the coefficient of variation as:

$$CV = \frac{\sqrt{ErrorMS}}{Grandmean} X100$$

Example: The experiment was aimed to determine the effect of three treatments (*T*1, *T*2 and *T*3) on average daily gain (g/d) of steers. Steers were weighed and assigned to four blocks according to initial weight. In each block there were three animals to which treatments were randomly assigned. Therefore, a total of 12 animals were used. Data with means and sums are shown in the following table:

	I	II	III	IV
T_1	826	865	795	850
T_2	827	872	721	860
T_3	753	804	737	822

Step 1: Computing the grand total, block total and treatment total

Blocks						
	I	II	III	IV	$\sum$ treatments	Treatment Means
T_1	826	865	795	850	3336	834
T_2	827	872	721	860	3280	820
T_3	753	804	737	822	3116	779
$\sum$ blocks	2406	2541	2253	2532	9732	
Block means	802	847	751	844		811

Step 2: Using t (treatments) and b (blocks), determine the degrees of freedom (d.f.) for each source of variation as follows:

1. D.f for Trt= t-1= 3-1=2
2. D.f for block=b-1=4-1=3
3. D.f for error=(t-1)(b-1)=(3-1)(4-1)=2x3=6
4. D.f for total =bxt-1=4x3-1=11

Step 3: Calculate the correction factor, and the various sums of squares (SS) as:

® Correction factor (C.F.) = $\frac{(GT)^2}{n} = \frac{(9732)^2}{12} = 7892652$

® Total SS = $\sum Y^2_{ij} - CF = \left[(826)^2 + ... + (822)^2\right] - 7892652 = 28406$

® Trt SS = $\frac{\sum_{i=1}^{t} T_i^2}{b} - CF = \frac{(3336)^2 + ... + (3116)^2}{4} - 7892652 = 6536$

® BSS= $\frac{\sum_{j=1}^{b} B_j^2}{t} - CF = \frac{(2406)^2 + ... + (2532)^2}{3} - 7892652 = 18198$

® E SS = Total SS – Treatment SS – block SS =28406-6536-18198=3672

Step 4: Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

- TRTMS = $\frac{TrSS}{t-1} = \frac{6536}{2}$ =3268
- Block MS = $\frac{BSS}{B-1} = \frac{18198}{3} = 6066$
- Error MS = $\frac{ESS}{(t-1)(b-1)} = \frac{3672}{6} = 612$

Step 5: Calculate the F value for testing significance of the treatment difference as:

$$(\text{TRT})\ F = \frac{TrtMS}{EMS} = \frac{3268}{612} = 5.34$$

$$(\text{Block})\ F = \frac{BMS}{EMS} = \frac{6066}{612} = 9.91$$

Step 6: Obtain the F values from the table, with f_1 = treatment *d.f* = (t-1) and f_2 = error *d.f* = (t-1) (b-1) by using given significance level (5% or 1%). For our example, the tabular F values with f1 = 2 and f2 = 6 degrees of freedom are 5.14 for the 5% level of significance and 10.92 for the 1% level.

Step 7: Construct ANOVA table and enter all the values computed from steps 2 to 6 and compare the computed F values of step 5 with the tabular F values of step 6, and decide on the significance among treatments using the following rules:

- ✓ If F Calculated > F tabulated at 1% ** (Highly significant)
- ✓ If F Calculated > F tabulated at 5% * (Significant)
- ✓ If F Calculated <= F tabulated at 5% ns (Non-significant)

The *ANOVA* table is:

Source	**Df**	**SS**	**MS**	**Computed F**	**Tabular F**		**Decision**
					5%	**1%**	
Block	3	18198	6066	9.91			
Treatment	2	6536	3268	5.34*	5.14	10.92	Reject H_0
Error	6	3672	612				
Total	11	28406					

☞ The critical value of *F* to teste the effect treatments with 2 and 6 degrees of freedom of treatment and error; and level of significance α = 0.05 is $F_{0.05}(2,6) = 5.14$. Since the calculated *F* = 5.34 is greater than the critical value at 5% level, H_0 is rejected indicating that significant differences exist among sample treatment means.

☞ Compute the efficiency of using randomized block design instead of completely randomized design. The efficiency of two experimental designs can be compared by calculating the relative efficiency (*RE*) of design 2 to design 1 (design 2 expected to have improvement in efficiency):

$$RE = \left(\frac{df_2 + 1}{(df_2 + 3)s_2^2} \right) \Big/ \left(\frac{df_1 + 1}{(df_1 + 3)s_1^2} \right)$$

Defining the completely randomized design as design 1, and the randomized block design as design 2; $s^2 1$ and $s^2 2$ are experimental error mean squares, and $df1$ and $df2$ are the error degrees of freedom for the completely randomized design and the randomized block design, respectively.

For the block design: $SSRES = 3672$; $s^2 2 = MSRES = 612$ and $df2 = 6$, $SSBLK = 18198$

For the completely randomized design:

$SS'RES = SSBLK + SSRES = 18198 + 3672 = 21870$

$df1 = 9$

$s^2 1 = SS'RES / df_2 = 21870 / 9 = 2430$

The relative efficiency is:

$$RE = \left(\frac{6+1}{(6+3)612}\right) \Big/ \left(\frac{9+1}{(9+3)2430}\right) = 3.71$$

☞ Since *RE* is much greater than one, the randomized block plan is better than the completely randomized design for this experiment.

2.3. Latin Square Design

2.3.1. Description

Sometimes an investigator may be aware of two sources of variability. In this case, the RCBD is no longer effective. The LSD is a simple extension of the RCBD that permits blocking in two directions. The LSD was originally conceived of for agriculture experiments to deal with gradients of moisture and soil nutrients. Nowadays there are many applications of this sort of design in biological experimentation.

☞ In the Latin square design treatments are assigned to blocks in two different ways, usually represented as columns and rows. Each column and each row are a complete block of all treatments. Hence, in a Latin square three explained sources of variability are defined: columns, rows and treatments. A particular treatment is assigned just once in each row and column. Often one of the blocks corresponds to animal and the other to period. Each animal will receive all treatment in different periods. The number of treatments (t) is equal to the number of columns and rows. The total number of measurements (observations) is equal to t^2.

2.3.2. Advantage and advantage of LSD

Advantage

- Not all rows and columns need to contain the same features of units.
- It reduces experimental error (Improve precision relative to CRD and RCBD)
  - ✓ This is because Latin square design controls two sources of variation (row and column variations) from the error term.

Disadvantage

- LSD 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.
- It is not flexible as compared to CRD and RCBD, because the assignment of treatment is depending of row and column
- It is difficult to add extra replications for certain treatments

2.3.3. Layout of the design

☞ If treatments are denoted with capital letters (A, B, C, D, etc.) then examples of 3 x 3 and 4 x 4 Latin squares are:

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

When an experiment is finished, the data computed as in the following table:

	1	2	3	4	R. total
1	C	D	B	A	r.1.
2	B	A	C	D	r.2.
3	D	C	A	B	r.3.
4	A	B	D	C	r.4.
C.total	c.1	c.2	c.3	c.4	G

Treatments	Treatment total
A	
B	
C	
D	

2.3.4. Analysis of Variance Model for LSD and Interpretation of Results

Model: $Y_{ijk} = \mu + r_i + c_j + t_k + e_{ijk}$

Where, μ = the overall mean

r_i = i^{th} row effect

c_j = j^{th} column effect

t_k = k^{th} treatment effect

e_{ijk} = the error tem

There are four sources of variation in a LS design; these are: row, column, treatment, and experimental error.

The results of LSD will be in the form of two-way table according to rows and columns. The results have to be arranged according to treatments also. Let R_i be the i^{th} row total, C_j be the j^{th} column total, The T_k be the K^{th} treatment total, and G.T. be the grand total. The different sum of squares for *t* x *t* LSD can be obtained as follows:

1. $C.F. = \frac{(G.T.)^2}{t^2}$

2. Total SS = $\sum X_{ijk}^{\ 2} - C.F.$

3. Row SS = $\frac{1}{t}\sum R_i^{\ 2} - C.F.$

4. Column SS = $\frac{1}{t}\sum C_j^{\ 2} - C.F.$

5. Treatment SS = $\frac{1}{t}\sum T_k^{\ 2} - C.F.$

6. Error SS = Total SS – Row SS - Column SS- Treatment SS

The result can be summarized in the form of analysis of variance (ANOVA) table.

Source	D.f	SS	MS	$F_{calculated}$	$F_{tabulated}$		Decision
					5%	1%	
Rows	t-1	RSS	RSS/t-1	RMS/EMS			
Column	t-1	CSS	CSS/t-1	CMS/EMS			
Treatments	t-1	TSS	TrSS/t-1	TrMS/EMS			
Error	(t-1)(t-2)	ESS	ESS/(t-1)(t-2)				
Total	T^2-1	TSS					

Example: Determine analysis of variance using data obtained on grain yield of three promising maize hybrids (A, B, and C) and a check (D) from an advanced yield trial with 4 x 4 Latin Square Design.

Row number	**Grain yield, t/ha**			
	Col. 1	Col. 2	Col. 3	Col. 4
1	1.640(B)	1.210(D)	1.425(C)	1.345(A)
2	1.475(C)	1.185(A)	1.400(D)	1.290(B)
3	1.670(A)	0.710(C)	1.665(B)	1.180(D)
4	1.565(D)	1.290(B)	1.655(A)	0.660(C)

Step 1: computing the column total, row total, grand total and treatment total

Row number	**Grain yield, t/ha**				**Row Total (R)**
	Col. 1	Col. 2	Col. 3	Col. 4	
1	1.640(B)	1.210(D)	1.425(C)	1.345(A)	5.620
2	1.475(C)	1.185(A)	1.400(D)	1.290(B)	5.350
3	1.670(A)	0.710(C)	1.665(B)	1.180(D)	5.225
4	1.565(D)	1.290(B)	1.655(A)	0.660(C)	5.170
Column total (C)	6.350	4.395	6.145	4.475	
Grand total (G)					**21.365**

Treatment	Total	Mean
A	5.855	1.464
B	5.855	1.471
C	4.270	1.068
D	5.355	1.339

Step 2: Using c (column), r (row), and t (treatment), determine the degrees of freedom (d.f.) for each source of variation as follows:

D.f for column=c-1=4-1=3
D.f for row = r-1= 4-1=3
D.f for treatment = t-1 = 4-1=3
D.f for error = (t-1) (t-2) = (4-1) (4-2) = 6
D.f for total= T^2-1 = 16-1= 15

Step 3: Calculate the correction factor, and the various sums of squares (SS) as:

C.F. = $\frac{(21.365)^2}{16} = 28.5289$

Total SS = $\sum X^2 - C.F. = [(1.640)^2 + (1.210)^2 + ... + (0.660)^2 - 28.5289 = 1.44139$

Row SS = $\frac{\sum R^2}{t} - C.F. = \frac{(5.620)^2 + (5.350)^2 + (5.225)^2 + (5.170)^2}{4} - 28.5289 = 0.0301$

Column SS = $\frac{\sum C^2}{t} - C.F. = \frac{(6.350)^2 + (4.395)^2 + (6.145)^2 + 4.475)^2}{4} - 28.5289 = 0.8273$

Treatment SS = $\frac{\sum T^2}{t} - C.F. = \frac{(5.855)^2 + (5.885)^2 + (4.270)^2 + (5.355)^2}{4} - 28.5289 = 0.4268$

ErrorSS=Total SS–Row SS–Column SS–Treatment SS = 1.4139-0.0301-0.8273-0.4268 = **0.1296**

Step 4: Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

Row Mean square = $\frac{RowSS}{t-1} = \frac{0.0301}{3} = 0.0100$

Column Mean square = $\frac{ColumnSS}{t-1} = \frac{0.8273}{3} = 0.2757$

Treatment Mean Square = $\frac{TreatmentSS}{t-1} = \frac{0.4268}{3} = 0.1422$

Error Mean square = $\frac{ErrorSS}{(t-1)(t-2)} = \frac{0.129585}{(3)(2)} = 0.021598$

Step 5: Calculate the F value for testing significance of the treatment difference as:

$$F = \frac{TreatmentMS}{ErrorMS} = \frac{0.142281}{0.021598} = 6.59$$

Step 6: Obtain the F values from the table, with f_1 = treatment *d.f* = (t-1) and f_2 = error *d.f* = (t-1) (t-2) by using given significance level (5% or 1%).

$F_{\text{tabulated } \alpha=5\% (3, 6 df) = 4.76}$

$\alpha=1\% (3, 6 df) = 9.78$

Step 7: Construct ANOVA table and enter all the values computed from steps 2 to 6 and compare the computed F values of step 5 with the tabular F values of step 6, and decide on the significance among treatments using the following rules:

- ✓ If F Calculated > F tabulated at 1% ** (Highly significant)
- ✓ If F Calculated > F tabulated at 5% * (Significant)
- ✓ If F Calculated <= F tabulated at 5% ns (Non-significant)

Source	Df	SS	MS	$F_{computed}$	$F_{tabular}$ 5%	1%
Row	3	0.030154	0.010051			
Column	3	0.827342	0.275781			
Treatment	3	0.426842	0.142281	6.59*	4.76	9.78
Error	6	0.129585	0.021598			
Total	15	1.413923				

Step 8: Conclusion: There is significant difference ($p<0.05$) on grain yield among four different varieties.

$$cv = \frac{\sqrt{ErrorMS}}{Grandmean} X100 = \frac{\sqrt{0.021598}}{1.33} X100 = 11.0\%$$

2.4. Split- Plot Design

2.4.1 Concepts of split plot design

The split-plot design is applicable when the effects of two factors are organized in the following manner. Experimental material is divided into several main units, to which the levels of the first factor are randomly assigned. Further, each of the main units is again divided into sub-units to which the levels of the second factor are randomly assigned.

In 2-factor experiments where it is desired to favor one factor over the other or where practical situation dictates application of one factor on larger plots, the split-plot design is used. In this design the main plot is being split into smaller subplots so as to accommodate a more precise measurement of the subplot factor and its interaction with the main plot factor. This means that the precision achieved by the subplot factor is at the expense of the precision for the measurement of the main plot factor.

For example, consider an experiment conducted on a meadow in which we wish to investigate the effects of three levels of nitrogen fertilizer and two grass mixtures on green mass yield. The experiment can be designed in a way that one block of land is divided into three plots, and on each plot a level of nitrogen is randomly assigned. Each of the plots is again divided into two subplots, and on each subplot within plots one of the two grass mixtures is sown, again randomly. The name split-plot came from this type of application in agricultural experiments. The main units were called plots, and the subunits split-plots. The split-plot design plan can include combinations of completely randomized designs, randomized block designs, or Latin square designs, which can be applied either on the plots or subplots.

2.4.2. Randomization and layout

Example: Suppose that we study the properties of 3 varieties of cotton breed for resistance to wilt as the subplot factor and 4 dates of sowing as the main plot factor, and the experiment is conducted in 4 replications.

The steps in randomization layout are given below.

1. Divide the experimental area into 4 blocks(r =4). Further divide each block into 4 main plots as shown in Figure 2.1.

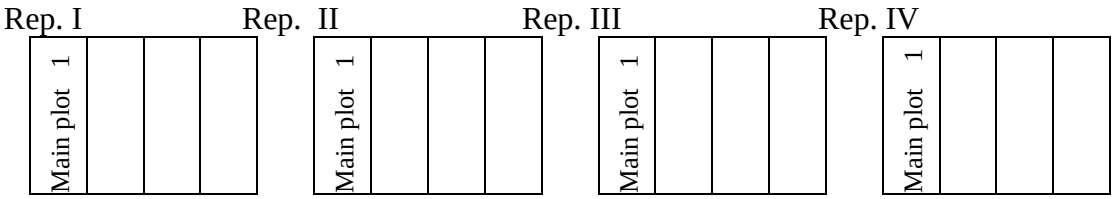

Figure 2.1 Layout of an experimental design into 4 blocks and 4 main plots

2. Randomly assign the main plot treatments Factor A, (4 dates of sowing: D_1, D_2, D_3, and D_4,) to the main plots among the 4 blocks or replications as shown in Figure 2.2

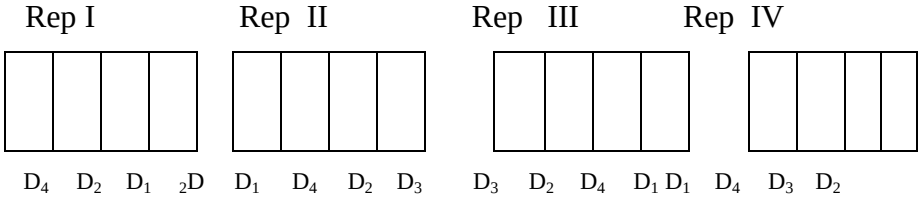

Figure 2.2. Random assignments of 4 sowing dates to each of the four main plots in each of the 4 replications

3. Divide each of the main plots in to 3 subplots and randomly assign the cotton varieties (Factor B) to each subplot as shown in Figure 2.3.

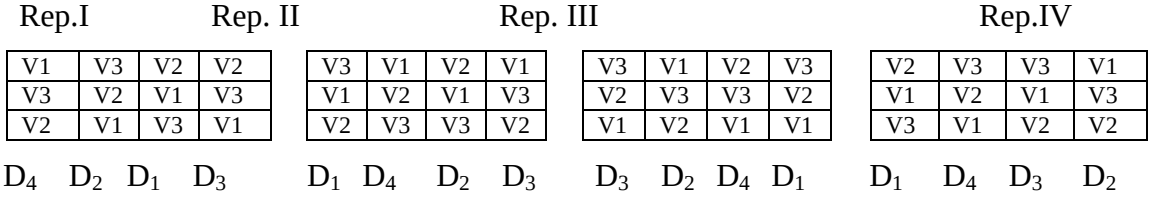

Figure 2.3. Random assignments of 4 subplot treatments to each of the 4 subplots in each replication.

We can see that the number of times the subplot factor is tested is much larger than that of the main plot factor. Therefore, the subplot factor is estimated more precisely than the main plot factor, and care should be taken when deciding on the factor to be assigned to the subplot. Also note that the major difference between RCBD and split plot lies on the concept of randomization. The dates of sowing are randomized on the main plots only, while the varieties are randomized on the sub-plots within main plot. In the case of RCBD, it is a factorial combination of date by variety which is randomized within a block. For example, since we have 16-treatment combination of date by variety, we normally randomize them on to 16 plots within each block.

In the case of split plot since we have two sized plots and randomization is being done at two stages, we have two different error terms.

2.4.3. ANOVA and Interpretation of Results

The model for this design is:

$y_{ijk} = \mu + R_k + A_i + \delta_{ik} + B_j + (AB)_{ij} + \varepsilon_{ijk}$

where: y_{ijk}= observation k in level i of factor A and level j of factor B

μ = the overall mean

R_k= the effect of the k^{th} of replication

A_i = the effect of level i^{th} of factor A

B_j= the effect of level j^{th} of factor B

$(AB)_{ij}$= the effect of the ij^{th} interaction of A x B

δ_{ik}= the main plot error

ε_{ijk}= the split-plot error

In performing the analysis of variance of such an experiment where the split-plot design is used, the experimenter must consider a separate analysis for the main plot factor (factor A) and the subplot factor (factor B) and The different sum of squares for SPD can be obtained as follows:

1. $C.F. = \dfrac{G^2}{rab}$
2. Total SS = $\sum x^2$ - C.F.
3. Replications(R) SS= $\dfrac{\sum R^2}{ab} - C.F.$
4. A (Trt) SS = $\dfrac{\sum A^2}{rb} - C.F.$
5. Error (A) SS = $\dfrac{\sum (AR)^2}{b} - C.F. - replication SS - ASS$
6. $B(subplot) SS = \dfrac{\sum B^2}{ra} - C.F.$
7. $AxB(main x \sup plot) SS = \dfrac{\sum (AB)^2}{r} - C.F. - BSS - ASS$
8. Error (b) SS = Total SS – (sum of all other SS)

The ANOVA table

Source	d.f	SS	MS	Fcal	Ftab		Decision
					5%	1%	
Replication	r-1	Rss	RMS				
A	a-1	SSA	AMS				
Error (a)	(r-1) (a-1)	Err SS	EaMS				
B	b-1	SSB	BMS				
B x A	(a-1)(b-1)	SSBA	ABMS				
Error (b)	a(r -1) (b -1)	SSE	EbMS				
Total	abr-1	TSS					

Example: Study carried out by agronomists to determine if differences in yield response to N fertilization exist among widely grown 5 hybrids of maize. For this study the main plot treatments were N rates of 0, 35, 70, and 105 kg/hectare and the subplot treatments were 5 hybrids (H_1, H_2, H_3, H_4, and H_5). The study was replicated 2 times. Grain yield data of 5 maize hybrids grown with 4 levels of Nitrogen in a split-plot experiment with 2 replications is given below.

		N rate Kg/ hectare **(factor A)**			
Replication (R)	Hybrid (**factor B**)	0	35	70	105
		Yield bu/ hectare			
I	H1	130	150	170	165
	H2	125	150	160	165
	H3	110	140	155	150
	H4	115	140	160	140
	H5	115	170	160	170
II	H1	135	170	190	185
	H2	150	160	180	200
	H3	135	155	165	175
	H4	130	150	175	170
	H5	145	180	195	200

Step 1: Computing the main and sub factor (treatment) total, replication total and grand total

	N rate/kg (factor A)				R.total
Rep	0	35	70	105	
1	595	750	805	790	2940
2	695	815	905	930	3345
T. Total	1290	1565	1710	1720	
G. total					6285

	N rate/kg (factor A)				H. Total
Hybrid (factor B)	0	35	70	105	
H1	265	320	360	350	1295
H2	275	310	340	365	1290
H3	245	295	320	325	1185
H4	245	290	335	310	1180
H5	260	350	355	370	1335
T. Total	1290	1565	1710	1720	
G. total					**6285**

Step 2: Using t (main treatment), h (sub treatment) r (replication), determine the degrees of freedom (D.F.) for each source of variation as follows:

1. Replication d.f. = r - 1 = 2 - 1 = 1
2. Main plot factor (A) d/f.= a- 1 = 4 - 1 = 3
3. Error (A) d.f. = (r - 1) (a - 1) = (2 - 1) (4 - 1) = 3
4. Subplot factor (B) d.f. = b -1 = 5 - 1 = 4
5. AXB d.f. = (a - 1)(b -1) =(4 - 1) (5 -1) = 12
6. Error (B) d.f. = a(r -1) (b -1) = 4(2 -1) (5 -1) = 16
7. Total d.f. = rab - 1 = 2 X 4 X 5 - 1 = 39

Step 3: Calculate the correction factor, and the various sums of squares (SS) as:

$$\text{Correctionfactor (C.F.)} = \frac{(GT)^2}{n} = \frac{(6285)^2}{40} = 987530.625$$

$$\text{Total SS} = \sum X^2{}_{ij} - C = \left[(130)^2 + (125)^2 + \ldots + (170)^2 + (200)^2\right] - 987530.625 = 20{,}244.38$$

$$\text{A SS} = \frac{\sum_{i=1}^{a} A_i{}^2}{bxr} - CF = \frac{(1290)^2 + (1565)^2 + (1710)^2 + (1720)^2}{5X2} - 987530.625 = 12{,}015.88$$

$$\text{RSS} = \frac{\sum R^2}{bxa} - CF = \frac{(2940)^2 + (3345)^2}{5x4} - 987530.625 = 4{,}100.63$$

$$\text{ESS of A} = \frac{\sum^{ra} RA^2}{b} - CF = \frac{(595)^2 + (695)^2 + ... + (930)^2}{5} - 987530.625 = 281.87$$

$$\text{BSS} = \frac{\sum^{b} B^2}{rxa} - CF = \frac{(1295)^2 + (1290)^2 + ... + (1335)^2}{2x4} - 987530.625 = 2466.25$$

$$\text{SS(BxA)} = \frac{\sum^{ab} BA^2}{r} - CF = \frac{(265)^2 + (320)^2 + ... + (355)^2 + (370)^2}{2} - Ass - Bss - 987530.625 = 863.75$$

E SS of (b) = TSS – RSS –ASS – ESS of A– BSS – SS (BxA)

=20,244.38 – 4,100.63 – 12,015.88 – 281.87 – 2,466.25 –863.75 = 479.97

Step 4: Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

1. $\text{TrtMS} = \frac{ASS}{a-1} = \frac{12,015.88}{3} = 4,017.29$
2. $\text{R MS} = \frac{RSS}{r-1} = \frac{4100.63}{1} = 4100.63$
3. $\text{EMS of A} = \frac{\text{ESS of A}}{(r-1)(a-1)} = \frac{281.87}{3} = 93.96$
4. $\text{BMS} = \frac{\text{SSB}}{b-1} = \frac{2,466.25}{4} = 616.56$
5. $\text{MS (BxA)} = \frac{\text{SS (BxA)}}{(b-1)(a-1)} = \frac{863.5}{12} = 71.98$
6. $\text{Error MS} = \frac{ESS}{a(r-1)(b-1)} = \frac{479.97}{16} = 29.99$

Step 5: Calculate the F value for testing significance of the treatment difference as:

$$F(\text{A}) = \frac{AMS}{EMS(A)} = \frac{12,015.88}{93.96} = 42.76 \qquad F(BxA) = \frac{MS(BxA)}{EMS(B)} = \frac{71.98}{29.99} = 2.40$$

$$F(B) = \frac{BMS}{EMS(B)} = \frac{616.56}{29.99} = 20.56$$

Step 6: Obtain the F values from the table,

1. f_1 = treatment *d.f* = (a-1) and f_2 = error *d.f* = (a-1) (r-1) by using given significance level (5% or 1%). The tabular F values with f1 = 3 and f2 = 3 degrees of freedom are 9.28 for the 5% level of significance and 29.46 for the 1% level.
2. f_1 = hybrid *d.f* = (b-1) and f_2 = error *d.f* = a(r -1) (b -1) by using given significance level (5% or 1%). The tabular F values with f1 = 4 and f2 = 16 degrees of freedom are 3.01 for the 5% level of significance and 4.77 for the 1% level.

3. f_1 = *d.f* of interaction of hybrid with nitrogen = (b-1) (a-1) and f_2 = error *d.f* = a(r -1) (b -1) by using given significance level (5% or 1%). The tabular F values with f1 = 12 and f2 = 16 degrees of freedom are 2.42 for the 5% level of significance and 3.55 for the 1% level.

Step 7: Construct ANOVA table and enter all the values computed from steps 2 to 6 and compare the computed F values of step 5 with the tabular F values of step 6, and decide on the significance among treatments using the following rules:

- ✓ If F Calculated > F tabulated at 1% ** (Highly significant)
- ✓ If F Calculated > F tabulated at 5% * (Significant)
- ✓ If F Calculated <= F tabulated at 5% ns (Non-significant)

The *ANOVA* table is:

Analysis of Variance of Hybrid X N response experiment in a split-plot design							
Source	d.f.	SS	MS	F-cal	F-tab		Decision
					5%	1%	
Total	39	20,244.38					
Replication	1	4,100.63	4,100.63				
Nitrogen(A)	3	12,015.88	4,017.29	42.76**	9.28	29.46	Accept H1
Error (a)	3	281.87	93.96				
Hybrid(B)	4	2,466.25	616.56	20.56**	3.01	4.77	Accept H1
Nitrogen X Hybrid (A X B)	12	863.75	71.98	2.40^{ns}	2.42	3.55	Reject H1
Error (b)	16	479.97	29.99				

Step 8: conclusion:

i) There was highly significant (p<0.01) difference on maize yield under different level of nitrogen.

ii) Similarly; there was highly significant (p<0.01) difference on maize yield among hybrids.

iii) But the interactions between nitrogen rate and hybrid showed no significance (p>0.05) on maize yield.

CHAPTER THREE: FACTORIAL EXPERIMENTS

At the end of the chapter student will able to:

- Understand the concept of factorial experiments
- Explain advantage and disadvantages of factorial experiments
- Explain source of variation of experimental unit for factorial experiment
- Understand the layout procedure of treatment on experimental unit
- Compare different treatment mean for factorial experiment

3.1.Concepts of Factorial Experiments

So far we have seen experiments for a single factor of main interest. Single Factor experiment means that the investigator is concerned in testing several levels of one factor while keeping all other factors at a constant level.

Factorial experiments simultaneously consider many factors of main interest; in addition, nuisance factors can be accommodated. An experiment, in which the levels of every factor are combined with the levels of every other factor in the same, experiment, is called a ***factorial experiment***.

A factorial experiment has two or more sets of treatments that are analyzed at *the same time*. Recall that treatments denote particular levels of an independent categorical variable, often called a **factor**. Therefore, if two or more factors are examined in an experiment, it is a factorial experiment. A characteristic of a factorial experiment is that all combinations of factor levels are tested. The effect of a factor alone is called a **main effect**. The effect of different factors acting together is called ***an interaction***.

* Consider an experiment to test the effect of protein content and type of feed on milk yield of dairy cows. The first factor is the protein content and the second is type of feed. Protein content is defined in **three levels**, and **two types** of feed are used. Each cow in the experiment receives one of the six proteins x feed combinations. This experiment is called a 3 x 2 factorial experiment, because three levels of the first factor and two levels of the second factor are defined. An objective could be to determine if cows' response to different protein levels is different with different feeds. This is the analysis of interaction. The main characteristic of a factorial experiment is the possibility to analyze interactions between

factor levels. Further, the factorial experiment is particularly useful when little is known about factors and all combinations have to be analyzed in order to conclude which combination is the best.

* Note that the term factorial describes a specific way in which the treatments are formed and does not, in any way, refer to the design used for laying out the experiment. For example, if the foregoing 2^2 factorial experiment is in a randomized block design, then the correct description of the experiment would be 2^2 *factorial experiment in a randomized complete block design*.
* The total number of treatments in a factorial experiment is the product of the number of levels of each factor; in the 2^2 factorial experiment, the number of treatments is 2 X 2 = 4, in the 2^3 factorial, the number of treatments is 2 X 2 X 2 = 8. The number of treatments increases rapidly with an increase in the number of factors or an increase in the levels in each factor. Indiscriminate use of factorial experiments has to be avoided because of their large size, complexity, and cost. Further it is not wise to commit oneself to a large experiment at the beginning of investigation when several small preliminary experiments may offer promising results.

3.2. Advantage and Disadvantages factorial experiments

Usually factorial experiments are more efficient than single-factor experiments but they can lead to an increase in complexity, size and cost of an experiment especially if some of the factors included are less important. Therefore, it is preferable to use an initial single factor experiment, especially if the number of potentially important factors is large.

Advantages of Factorial Arrangement

1. Provides estimates of interactions (The advantage of factorial structure, over single factor experiment, is that it will enable us to simultaneously test effects of several factors and their cross effects known as ***interactions.*** Interaction provides a measure of the pattern of the effect of one factor over the other factor).
2. Possible increase in precision due to so-called" hidden replication"
3. Experimental rates can be applied over a wide range of conditions

Disadvantages of Factorial Arrangement

1. Some treatment combinations may be of little interest
2. Experimental error may become large with a large number of treatments
3. Interpretation may be difficult (especially for 3-way or more interactions)

3.3. Layout of factorial experiments

⇒ **Factor**: refers to a kind of treatment. Factors will be referred to with capital letters.

⇒ **Level**: refers to several treatments within any factor. Levels will be referred to with lower case letters. A combination of lower case letters and subscript numbers will be used to designate individual treatments (a_0, a_1, b_o, b_1, a_0b_0, a_0b_1, etc)

⇒ **Interaction**: (1) the failure for the response of treatments of a factor to be the same for each level of another factor. (2) When the **simple effects** of a factor differ by more than can be attributed to chance, the differential response is called an **interaction**.

♦ When the factorial experiment has been chosen, we know the total number of treatment combinations that must be tested. For this number of treatment design such as CRD, RCBD, or LSD may be adopted. For example for a 3^2 factorial experiment there will be 9 treatment combinations. We can arrange the 9 treatments in RCBD or 9 X 9 LSD. Depending upon the basic design the randomization is carried out. Thus, we can have factorial CRD, factorial RCBD and factorial LSD. Relative advantages of these designs are the same as that of simple CRD, RCBD, and LSD.

♦ **Randomizing factorial Arrangement**

1. Assign numbers to treatment combinations
2. Randomize treatments according to deign
3. Example RCBD with a 2 X 4 Factorial Arrangement

Treatment	Treatment number	Treatment	Treatment number
a_0b_0	1	a_1b_0	5
a_0b_1	2	a_1b_1	6
a_0b_2	3	a_1b_2	7
a_0b_3	4	a_1b_3	8

Rep I	3 a_0b_2	7 a_1b_2	2 a_0b_1	6 a_1b_1	4 a_0b_3	5 a_1b_0	1 a_0b_0	8 a_1b_3

3.4. Types of Factorial Experiments

There can be *two, three, or more* factors in an experiment. Accordingly, factorial experiments are defined by the number, two, three, etc., of factors in the experiment.

The Two Factor Factorial Experiment: Consider a factorial experiment with two factors *A* and *B*. Factor *A* has *a* levels, and factor *B* has *b* levels. Let the number of experimental units for each *A*x *B* combination be *n*. There is a total of *nab* experimental units divided into *ab* combinations of *A* and *B*. The set of treatments consists of *ab* possible combinations of factor levels.

The **model** for a factorial experiment with two factors *A* and *B* is:

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \qquad i = 1,\ldots,a; \quad j = 1,\ldots,b; \quad k = 1,\ldots,n$$

where:

- y_{ijk}= observation *k* in level *i* of factor *A* and level *j* of factor *B*
- μ = the overall mean
- A_i = the effect of level *i* of factor *A*
- B_j= the effect of level *j* of factor *B*
- $(AB)_{ij}$= the effect of the interaction of level *i* of factor *A* with level *j* of factor *B*
- ε_{ijk}= random error with mean 0 and variance $\sigma 2$

  a = number of levels of factor *A*; *b* = number of levels of factor *B*; *n* = number of observations for each *A x B* combination.

3.4. Calculation of Simple effects, Main effects and Interactions

- Simple effects, main effects and interaction will be explained using the following data set:

Table 1.Effect of two nitrogen rates of fertilizer on grain yield (kg/ha) of two barley cultivars.

	Fertilization	
Cultivar (A)	0 kg/ha (b_0)	60 kg/ha (b_1)
Laker (a_0)	1.0 (a0b0)	3.0 (a0b1)
Morex (a_1)	2.0 (a1b0)	4.0 (a1b1)

- The **simple effect** of a factor is the difference between its two levels at a given level of the other factor.

⇒ Simple effect of A at b0 = a1b0 - a0b0 = 2-1 = 1

⇒ Simple effect of A at b1 = a1b1 – a0b1 = 4-3 = 1

⇒ Simple effect of B at a0 = a0b1 – a0b0 = 3- 1 = 2

⇒ Simple effect of B at a1 = a1b1 – a1b0 = 4 -2 = 2

- ❖ The **main effect** of a factor is the average of the simple effects of that factor over all levels of the other factor.

  ⇒ Main effect of A = (Simple effect A at a0 + simple effect of A at b1)/ 2 = (1+1)/2 = 1

  ⇒ Main effect of B = (Simple effect of B at a0 + simple effect of B at a1)/2 = (2+2)/2 = 2

- ❖ The interaction is a function of the difference between the simple effects of A at two levels of B divided by two, or vice-versa. This works only for 2 x 2 factorial

  ⇒ AXB = ½ (Simple effects of A at b1 – Simple effects of A at b0) =1/2 (1-1) = 0

  ⇒ A X B = ½ (Simple effects of B at a1 – Simple effects of B at a0) = ½ (2-2) = 0

Some facts of interactions:

a. An interaction between two factors can be measure only if the two factors are tested together in the same experiment.
b. When an interaction is absent, the simple effect of a factor is the same for all levels of the other factors and equals the main effect.
c. When the interactions are present, the simple effect of a factor changes as the level of the other factor changes. Therefore, the main effect is different from the simple effects.

Example of analysis of variance for 2 X 2 factorial experiments

Assume an experiment was conducted to determine the effect of species difference and temperature difference on growth rate of fish. Note: species difference is denoted by factor A with 2 species (a_1 and a_2) and temperature by factor B with two levels (b_1 and b_2)

Replicate	Treatment				
	a1b1	a1b2	a2b1	a2b2	$\mathbf{Y_{ij}}$
1	12	19	29	32	**92**
2	15	22	27	35	**99**
3	14	23	33	38	**108**
4	13	21	30	37	**101**
Yi.	**54**	**85**	**119**	**142**	**400=Y…**

A	B		A total
	b1	b2	
a1	54	85	139
a2	119	142	261
B total	173	227	400=grand total

Step 2: Using a and b (main effect of factor A and B) and r (replications), determine the degrees of freedom (D.F.) for each source of variation as follows:

1. Replication d.f. = r - 1 = 4 - 1 = 3
2. factor (A) d/f.= a- 1 = 2 - 1 = 1
3. factor (B) d.f. = b -1 = 2 - 1 = 1
4. AX B d.f. = (a - 1)(b -1) =(2 - 1) (2 -1) = 1
5. Error d.f. = (r -1) (ab -1) = (4 -1) (2x2 -1) =9
6. Total d.f. = rab - 1 = 4X 2 X 2 - 1 = 15

Step 3. Calculate Correction factor and sum square of different source variation.

$$1.CF = \frac{Y^2}{rab} = \frac{400^2}{4x2x2} = 10{,}000$$

$$2.TotalSS = \sum Y_{ij}^{\ 2} - CF = (12^2 + 15^2 + 14^2 + ... + 37^2) - CF = 1{,}170$$

$$3.\text{Re}\,pSS = \frac{\sum Y_j^{\ 2}}{ab} - CF = \frac{92^2 + 99^2 + 108^2 + 102^2}{2x2} - CF = 32.5$$

$$4.ASS = \frac{\sum A^2}{rb} - CF = \frac{(139^2 + 261^2)}{4x2} - CF = 930.25$$

$$5.\text{B SS} = \frac{\sum B^2}{ra} - CF = \frac{(173^2 + 227^2)}{4x2} - CF = 182.25$$

$$6.\text{A X B SS} = \frac{\sum ab^2}{r} - ASS - BSS - CF = \frac{(54^2 + 85^2 + 119^2 + 142^2)}{4} - 930.25 - 182.25 - 10000 = 4$$

Note: A SS + B SS +A X B SS = Treatment SS

7. Error SS = Total SS – Rep SS – A SS – B SS – A X B SS = 21

Step 4: Calculate the mean square (MS) for each source of variation by dividing each SS by its corresponding d.f.

$$1.\ \text{AMS} = \frac{ASS}{a-1} = \frac{930.3}{1} = 930.3$$

$$2.\ \text{R MS} = \frac{RSS}{r-1} = \frac{32.5}{3} = 10.83$$

$$3.\ \text{BMS} = \frac{\text{SSB}}{b-1} = \frac{182.3}{1} = 182.3$$

$$4.\ \text{MS (BXA)} = \frac{\text{SS (BXA)}}{(b-1)(a-1)} = \frac{4}{1} = 4$$

$$5.\ \text{Error MS} = \frac{ESS}{\text{(r-1)(ab-1)}} = \frac{21}{9} = 2.33$$

Step 5: Calculate the F calculated (F.cal) value for testing significance of the treatment difference as follows:

$$1.F(\text{A}) = \frac{AMS}{EMS} = \frac{930.3}{2.33} = 398.7$$

$$2.F(B) = \frac{BMS}{EMS} = \frac{182.3}{2.33} = 78.11$$

$$3.F(BXA) = \frac{MS(BXA)}{EMS} = \frac{4.00}{2.33} = 1.71$$

Step 6: Obtain the F values from the table,

1. f_1 = factor A *D.f* = (a-1) and f_2 = error *D.f* = (r-1) (ab-1) by using given significance level (5% or 1%). The tabular F values with f1 = 1 and f2 = 9 degrees of freedom are **5.12** for the 5% level of significance and **10.6** for the 1% level.
2. f_1 = factor B *D.f* = (b-1) and f_2 = error *D.f* = (r-1) (ab-1) by using given significance level (5% or 1%). The tabular F values with f1 = 1 and f2 = 9 degrees of freedom are **5.12** for the 5% level of significance and **10.6** for the 1% level.
3. f_1 = *D.f* of interaction of A with B= (a-1) (b-1) and f_2 = error *D.f* = (r-1) (ab-1) by using given significance level (5% or 1%). The tabular F values with f1 = 1 and f2 = 9 degrees of freedom are **5.12** for the 5% level of significance and **10.6** for the 1% level.

Step 7. Construct ANOVA table and enter all the values computed in steps 3 to 6 and Compare the computed F values of step 5 with the tabular F values of step 6, and decide to reject Ho if $\text{Fcal} \geq F_{tab}$ or accept Ho if $\text{Fcal} \leq F_{tab}$.

Source	D.f	SS	MS	F.cal	F. tab		Decision
					0.05	0.01	
Rep	r-1=3	32.5	10.83				
A	a-1=1	930.3	930.3	398.7**	5.12	10.6	Accept H1
B	b-1=1	182.3	182.3	78.11**	5.12	10.6	Accept H1
AXB	(a-1)(b-1) = 1	4	4.00	1.71^{ns}	5.12	10.6	Accept H0
Error	(r-1)(ab-1) = 9	21	2.33				
Total	rab-1 = 15	1170					

Step 8: Conclusion

I. *There was highly significant (p<0.01) difference between different species of fish on growth rate fish.*

II. *Similarly; There was highly significant (p<0.01) difference between the two level of temperature on growth rate of fish.*

III. *But the interactions between species difference and temperature difference has no a significance (p>0.05) difference on growth rate of fish*

Coefficient of variation of factorial design

$$Grandmean = \frac{grandtotal}{rab} = \frac{400}{4x2x2} = 25$$

$$cv = \frac{\sqrt{ErrorMS}}{Grandmean} X100 = \frac{\sqrt{2.33}}{25} X100 = 6.11\%$$

Since the CV is much smaller than the optimum, so the reliability of the experiment is high.

3.7. Interpreting Results of ANOVA involving interaction terms

- Interpretation should always begin with the higher level interaction terms (e.g. three-way interactions before two-interactions, etc.).
- Interpretation of the main effects should never be done before interpreting the interaction terms.
- The F-test for interaction terms can be significant because of two reasons.
  i. True interaction
  ii. Differences in magnitude between treatment

CHAPTER FOUR: COMPARISON BETWEEN TREATMENT MEANS

At the end of the chapter student will able to:

- Understand the concept of "f" Test, p-value and types of error
- Describe the concept of paired comparison
- Describe the concept of group comparison

4.1. The "f" Test and Analysis of Variance (ANOVA)

A statistical test can have only two results: to reject or fail to reject (accept) the null hypothesis H_0. Consequently, based on sample observations, there are two possible errors:

- type I error = reject of H_0, when H_0 is actually true
- type II error = failure to reject H_0, when H_0 is actually false

An F test is used to conclude; is there any significant difference among groups or treatments. If H_0 is **not rejected (accept),** it is not necessary or appropriate to further analyze the problem, although the researcher must be aware of the possibility of a type II error. When H_0 is **rejected**, it is appropriate to further question which treatment(s) caused the effect, that is, between which groups is the significant difference found. If there are more than two means, it is not advisable to use the t-test. When **several means** are compared analysis of variance is the most powerful test. The t-test is good if there are one or two means. The analysis of variance is the powerful statistical technique developed for analyzing measurements that depend on several kinds of effects which operate simultaneously, to decide which kind of effects are important and also to estimate these effects. It is a powerful technique, which allows analysis and interpretation of observations from several populations. This versatile statistical tool partitions the total variation in a data set according to the **source of variation** that is present.

The ANOVA for two groups is identical to the results obtained with a **t test**; it is fair to say that ANOVA is an extension of the t test to handle more than two independent groups. The F test can be used to compare two sample variances to determine whether they are equal. It can also be used to compare three or more means. When three or more means are compared, the technique is called analysis of variance (ANOVA). The ANOVA technique uses two estimates of a population variance. For one independent variable, the analysis of variance is called a one-way

ANOVA. When there are two independent variables, the analysis of variance is called a two-way ANOVA. The theoretical basis for performing the ANOVA test is the partitioning of the available variance of all observations into two sources of variations- variation between the group means and variation within each of the groups. The sampling distribution used for testing these means is not the t distribution but rather the F distribution.

The F-test indicates significant differences among the treatment means tested. However, it does not identify the specific pairs of treatment means that differed. F-test is not able to answer the question of whether every one of the treatment means gave significantly higher yields than that of the other treatment. To answer this question the procedure for mean comparison will be discussed later.

What is a P-value?

When you perform a hypothesis test in statistics, a *p*-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the *null hypothesis.*

The *alternative hypothesis* is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a *p*-value to weigh the strength of the evidence (what the data are telling you about the population).

The *p*-value is a number between 0 and 1 and interpreted in the following way:

- A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- *P-values* very close to (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

The alternative hypothesis is what we expect to be true if the null hypothesis is false. We cannot prove that the alternative hypothesis is true but we may be able to demonstrate that the alternative is much more reasonable than the null hypothesis given the data. This demonstration is usually expressed in terms of a probability (a P-value) quantifying the strength of the evidence against the null hypothesis in favor of the alternative.

P > 0.10	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
0.05 < P < 0.10	Weak evidence against the null hypothesis in favor of the alternative.
0.01 < P < 0.05	Moderate evidence against the null hypothesis in favor of the alternative.
0.001 < P < 0.01	Strong evidence against the null hypothesis in favor of the alternative.
P < 0.001	Very strong evidence against the null hypothesis in favor of the alternative.

Comparison between Treatment Means

There are many ways to compare the means of treatments tested in an experiment. Only those comparisons helpful in answering the experimental objectives should be thoroughly examined. **Consider** an experiment in rice weed control with 15 treatments, four hand weeding, ten herbicides and 1 with no weeding (control). The probable questions that may be raised and the specific mean comparisons that can provide their answers may be:

- ☞ Is any treatment effective in controlling weeds? This could be answered simply by comparing the means of the non-weeded treatment with the mean of each of the 14 weed–control treatments.
- ☞ Are there differences between the 14 weed-control treatments? If so, which is effective and which is not? Among the effective treatments, are their differences in levels of affectivity? If so, which is best? To answer this question, the means each of the 14 weed control treatments is compared to the control's mean and those that are significantly better than the control are selected. In addition, the selected treatments are compared to identify the best among them.
- ☞ Is there any difference between the group of hand weeding treatments and the group of herbicide treatments? To answer this question, the means of the four hand weeding treatments are averaged and compared with the averaged means of the 10 herbicides treatments.
- ☞ Are there differences between the four hand weeding treatments? If so, which treatment is best? To answer these questions, the four hand-weeding treatment means are compared to detect any significant difference among them and the best treatments are identified.

These different types can, however, be classified either as **pair comparison or group comparison**.

Pair comparison

Pair comparison is the simplest and most commonly used comparison in agricultural research. There are many procedures for pair-wise comparisons of means. Such as the Least Significance Difference, Tukey tests, Bonferoni, Newman-Keuls, Duncan, Dunnet, etc.

There are two types of Pair comparison:

A. **Planned pair comparison**, in which the specific pair of treatments to be compared was identified before the start of the experiment. A common example is comparison of 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 is compared to identify pairs of treatments that are significantly different.

The two most commonly used test procedures for pair comparisons in agricultural research are the **least significant difference (LSD) test** which is suited for a **planned pair comparison** and **Duncan's multiple range test (DMRT)** which is applicable to an **unplanned pair comparison**.

The Least Significance Difference Method

Least significance difference (LSD) is used to test pair wise comparisons only if the null hypothesis is rejected in the analysis of variance **F** test. If the null hypothesis is not rejected on the basis of F test, then all treatments means are assumed to be the same and no further testing is done. An **advantage** of the *LSD* is that it has a low level of ***type II error*** and will most likely detect a difference if a difference really exists. A disadvantage of this procedure is that it has a high level of ***type I error***. Because of the probability of type I error, a significant *F* test must precede the *LSD* in order to ensure a level of significance α for any number of comparisons.

Duncan's Multiple Range Test (DMRT)

Drawbacks of Duncan's

It requires a greater observed difference to detect significance. Therefore, duncans is termed as conservative, because some differences which may not be significant in Duncans could be significant in other tests. It needs multi-valued critical value so that the difference between treatments required for significance depends on the number of treatments in experiment. In reality however, it is hard to believe that the difference between two treatment means depends in any way on what others were included in the experiment.

Tukey Test

An advantage of the Tukey test is that it has fewer incorrect conclusions of $\mu i \neq \mu i$' (type I errors) compared to the *LSD*; a disadvantage is that there are more incorrect $\mu i = \mu i$' conclusions (type II errors).

Group comparison

For group comparison, more than two treatments are involved in each comparison. There are four types of comparison:

i. **Between-group comparison,** in which treatments are classified in to s (where s>2) meaningful groups, each group consisting of one or more treatments, and the aggregate mean of each group is compared to that of the others.

ii. **Within-group comparison,** which is primarily designed to compare treatments belonging to a subset of all the treatments tested. This subset generally corresponds to a group of treatments used in the between group comparison. In some instances, the subset of treatments in which the within group comparison is to be made may be selected independently of the between group comparison.

iii. **Trend comparison**, which is designed to examine the functional relationship between treatments levels and treatment means. Consequently, it is applicable only to treatments that are quantitative.

iv. **Factorial comparison**, which as the name implies, is applicable only to factorial treatments in which specific sets of treatment means are compared to investigate the main effects of the factors tested and , in particular, the nature of their interaction.

CHAPTER FIVE: PROBLEM DATA

At the end of the chapter student will able to:

- Understand the concept of problem data.
- Describe what missing data is?
- Mention and describe the cause of missing data.

Problem Data

Analysis of variance is valid for once research only if the basic research data satisfy certain conditions. Some conditions are implied, others are specified. In field experiments, for example, it is **implied** that all plots are grown successfully and all necessary data are taken and recorded. In addition, it is **specified** means that the data satisfy all the mathematical assumptions underlying the ANOVA. We use the term **problem data** for any set of data that does not satisfy the implied or the stated conditions for valid analysis of variance. **Two groups of problem data** that are commonly encountered in agricultural research is **missing data** and **data that violate some assumptions of the analysis of variance**

A. Missing data

A missing data situation occurs whenever a valid observation is not available for any one of the experimental units. Even though data gathering in field experiments is usually done with extreme care, numerous factors beyond the researchers control can contribute to missing data.

Common causes of missing data:

1. Improper treatment
2. Destruction/loss of experimental material
3. Loss of harvested/collected samples
4. Illogical data

1. Improper treatment

Improper treatment is declared when an experiment has one or more experimental plots/materials that do not receive the intended treatment. No applications, applications of incorrect dose, and wrong timing of applications are common cause of improper treatment. Any observation made on an experimental material/plot where treatment has not been properly applied should be considered invalid.

2. Destruction/loss of experimental material

Most field experiments aim for a perfect stand in all experimental plots/materials but that is not always achieved. This may occur due to loss of animals, poor germination, and cause of disease for the experimental materials. The destruction/loss of the experimental animals/plants/materials must not be the result of the treatment effect.

3. Loss of harvested/collected samples

Many animal/plant characters cannot be conveniently recorded, either in the field or immediately after harvest. Harvested/collected samples may require additional processing before the required data can be measured. For example, milk composition analysis, grain yield of rice. Some characters may involve long sampling and measurement processes or may require specialized and elaborate measuring devices. Example: protein contents are measured in laboratory, breeding data.

4. Illogical data

In contrast to the case of missing data where the problem is recognized ***before data are recorded***, illogical data are usually recognized ***after the data*** have been recorded and transcribed. Data may be considered illogical if their values are too extreme to be considered ***within the logical ranges of the normal behavior*** of the experimental materials. However, only illogical data resulting from some kind of error can be considered as missing. Common errors resulting in illogical data are misread observations, incorrect transcription and improper applications of the sampling technique or the measuring instrument.

B. Data that violate the Assumptions of ANOVA

The usual interpretation of the analysis of variance is valid when certain mathematical assumption concerning the data is met. These assumptions are:

- **Additive effects.** Treatment effects and environmental effects are additive.
- **Independence of errors.** Experimental errors are independent.
- **Homogeneity of variance.** Experimental errors have common variance.
- **Normal distribution**. Experimental errors are normal distributed**.**

Failure to meet one or more of these assumptions is called **violation of assumption of ANOVA**.

CHAPTER SIX: CHECKING THE ASSUMPTIONS AND TRANSFORMATION OF DATA

At the end of the chapter student will able to:

- Explain the concept of assumption of ANOVA
- Explain the concept of transformation of data
- Mention and describe data transformation techniques

6.1. The Assumption of ANOVA

The theoretical concepts of analysis of variance are based on a set of assumptions. The use of ANOVA test we have to be made the following assumption:

Randomization: All ANOVAs require that sampling of individuals be at ***random***. *Adequate safeguards to ensure random sampling during the design of experiment, or when samplings from natural populations are essential.*

Independence: The second assumption is that the items be ***independent***. *This may not be true if there is correlation in time or space.*

Equality of variances: Equality of variances in a group of samples is an important precondition for several statistical tests.

Normality of distribution: each group sample is drawn from a normal distributed population.

Additivity of effects: The **additivity** effects in a 2-way or higher order ANOVA should be *small.*

- *If the above assumptions are **not fulfilled**, you should carry out one of several possible alternative steps to remedy the situation.*

Departure from the assumptions of analysis of variance can be rectified by "***transformation***" of the original data by use of new scale.

6.2. Data Transformation Techniques

Data transformation is the most appropriate remedial measure for variance heterogeneity where the variance and the mean are functionally related. With this technique, the original data are converted in to a new scale resulting in a new data set that is expected to satisfy the condition of homogeneity of variance. Because of a common transformation scale is applied to all observations, the comparative values between treatments are not altered and comparisons

between them remain valid. The appropriate data transformation to be used depends on the specific type of relationship between the variance and the mean.

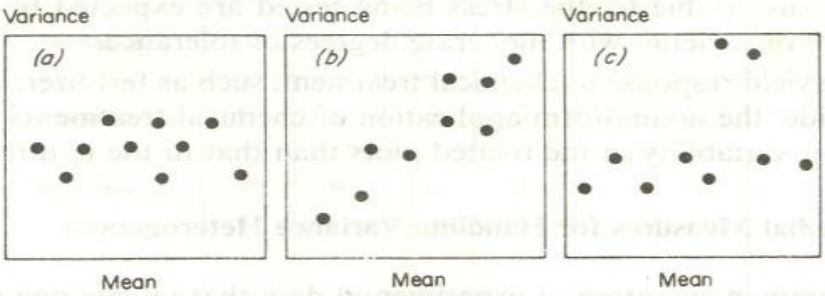

Figure 7.2 Illustration of different types of variance heterogeneity: (*a*) homogeneous variance, (*b*) heterogeneous variance where variance is proportional to the mean, and (*c*) heterogeneous variance without any functional relationship between variance and mean.

- Three of the most commonly used transformations for data in agricultural research:
    I. **Logarithmic Transformation**
    II. **Square-root Transformation**
    III. **Arc Sine Transformation**

I. Logarithmic Transformation

Logarithmic transformation is most appropriate for data where the standard deviation is proportional to the mean or where the effects are multiplicative. These conditions are generally found in data that are whole numbers and cover a wide range of values.

Data having an exceedingly wide range

Either count or measurement data which has extremely wide range would likely has the variances proportional to the square of the treatment means. For example if the range of data of three treatment means are, A, 3 to 18; B, 360 to 950; and c, 8000 to 15,000 a logarithmic transformation will be appropriate. When zero values are encountered and when some of the values are less than 1.0 the transformation will be log (X+1).Following the analysis the means can be transformed back to the original scale using an **antilog table**.

II. Square-Root Transformation

It is appropriate for data consisting small whole numbers and are limited to values of 10 or less.

- The suggested correction for this is to transform the data by using. $\sqrt{X}$
- If any zeros are encountered the transformation might be. $\sqrt{X+0.5}$

Analysis of the transformed data will be handled by the usual methods. Following the analysis the mean values will be transformed back to the raw data. It is also appropriate for percentage data where the range is between 0 and 30% or between 70 and 100%

6.2.3. Arc Sine Transformation

Percentage data

If the data to be analyzed are recorded as percentages and the range of the data is between 30 and 70 then most authors consider the need for transformation doubtful. However, if the values do not fall within this range the size of the standard error will be related to the size of the percentage.

- The angle $=\sqrt{perentage}$; arcsine can be obtained directly from the table statistical table.

Again the analysis is performed using the transformed values and the means are transformed back to the raw values. Not all percentage data need to be transformed and, even if the data need to be translated, arc sine transformation technique is not the only transformation technique. The following rules are may be useful in choosing the proper transformation scale for percentage data derived from count data.

- For percentage data lying within the range of 30 to 70%, no transformation is needed.
- For percentage data lying within the range of either of 0 to 30%, or 70 to 100%, but not both, the square-root transformation should be used.
- For percentage data that do not follow the ranges specified in either rule 1 and 2, the arc sine transformation should be used.

CHAPTER SEVEN: SIMPLE LINEAR CORRELATION AND REGRESSION

At the end of the chapter student will able to:

- Explain the concept of simple linear correlation
- Explain the concept of simple linear regression

7.1. Simple Linear Correlation

The two most common methods used to describe relationships between two quantitative variables (X and Y) are **linear correlation** and **linear regression**. The purpose of correlation analysis is to measure and interpret the **strength of a linear relationship** between **two continuous** variables. When conducting correlation analysis, we use the term **association** to mean "***linear association***". The Pearson correlation coefficient is also known as the ***simple correlation coefficient*** **(r)**, *product movement correlation coefficient*, or *coefficient of correlation (there are other types of correlation coefficients)*. The coefficient of correlation measures the strength of the linear relationship between two variables.

Scatter Diagram

- There are three types of correlations to display in **scatter Diagram**:
  - (1) Perfectly correlated;
  - (2) Partially correlated and
  - (3) Uncorrelated.

The correlation is used when there is interest in determining the degree of association among variables, but when they cannot be easily defined as dependent or independent. For example, we may wish to determine the relationship between weight and height, but do not consider one to be dependent on the other, perhaps both are dependent on a third factor.

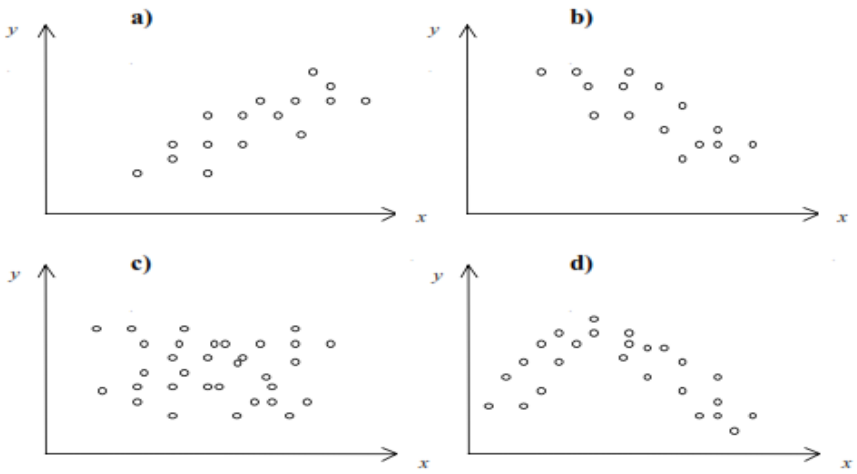

Figure 8.1 a) positive correlation, b) negative correlation, c) no association, and d) an association but it is not linear

7.1.1. Correlation coefficient

Correlation coefficient measures, the strength of the degree of association between two random variables. The value of r can range from -1 to +1, and is independent of units of measurements and when r=0, the two variables are not correlated. For independent random variables, the value is close to zero, while it is positive for directly related variables and negative for those that are inversely related. The higher the absolute value of r, the closer the degree of association. The strength of the association increases as r approaches to the absolute value of 1.0.The coefficient of correlation is estimated from a random sample by a coefficient of correlation (r):

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

Where: r= stands for correlation coefficient

X and Y = stands for the value the variable

n = stands for sample size

Interpretation of correlation coefficient

Correlation value	Direction & strength of correlation
-1.0	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakley negative
0.0	No correlation
+0.2	Weakly positive
+0.5	Moderately positive
+0.8	Strongly positive
+1.0	Perfectly positive

Calculating of correlation coefficient

EX: Estimate the correlation coefficient of milk yield (Y) and feed intake (X) of five cows:

X	76	75	70	66	78
Y	15	14.8	13.8	13	15.5

To calculate r, we have to calculate the sum X, Y, XY and sum square of x and Y

Ser. No.	Y	X	Y^2	X^2	XY
1	15	76	225	5776	1140
2	14.8	75	219.04	5625	1110
3	13.8	70	190.44	4900	966
4	13	66	169	4356	858
5	15.5	78	240.25	6084	1209
$\sum$	72.1	365	1043.73	26741	5283

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

$$r = \frac{5283 - \frac{365x72.1}{5}}{\sqrt{[26741 - \frac{(365)^2}{5}][1043.73 - \frac{(72.1)^2}{5}}}$$

$$r = \frac{5283 - 5263.3}{\sqrt{[26741 - 26645][1043.73 - 1039.682}}$$

$$r = \frac{19.7}{\sqrt{[96][4.048}}$$

$$r = \frac{19.7}{\sqrt{388.608}} \qquad r = \frac{19.7}{19.71} = 0.99$$

There for the correlation coefficient between these two variables are 0.97 and this indicates the relationship between weight of hen and feed intake is strong positive correlation.

7.2. Simple Linear Regression

It is often of interest to determine how changes of values of some variables influence the change of values of other variables. For example,

- How alteration of air temperature affects feed intake, or
- How increasing the protein level in a feed affects daily gain.

In both the first and the second above examples, the relationship between variables can be described with a function, a function of temperature to describe feed intake, or a function of protein level to describe daily gain. A function that explains such relationships is called a **regression function** and analysis of such problems and estimation of the regression function is called **regression analysis.** Regression includes a set of procedures designed to study statistical relationships among variables in a way in which one variable is defined as **dependent** upon others defined as **independent variables**.

By using regression the **cause-consequence relationship** between the independent and dependent variables can be determined. In the examples above, feed intake and daily gain are dependent variables, and temperature and protein level are independent variables. Recall that the main goal of regression analysis is to determine the functional dependency of the dependent variable *y on the independent variable x.*

The dependent variable is usually denoted by *y and the* independent variables by *x.* The roles of both variables x and y are clearly defined as dependent and independent. The values of y are expressed as a function of the values of x. *Often the dependent variable is also called the* ***response*** **variable**, and the independent variables are called **repressors or predictors**. Of two variables under study one may represent the **cause** and the other may represent the **effect**. The variable representing the cause is known as ***independent*** **(*input*)** variable (X) or ***predictor*** **variable or *repressor***. The variable representing the effect is known as ***dependent*** **variable (Y)** or sometimes called the ***predicted*** **variable or *outcome*** **variable.**

If the relationship between the two variables is a straight line, it is known as *simple linear regression*; otherwise it is called simple non-linear regression. **Multiple regression** procedures are utilized when the change of a dependent variable is explained by changes of **two or more independent variables**. A multiple regression is regression that has two or more independent variables. For example, weight gain in animal may be influenced by the protein level in feed, the amount of feed consumed, and the environmental temperature, etc. The variability of a dependent variable y can be explained by a function of several independent variables, x_1, $x_2 \ldots x_p$.

Purpose of regression:

1. To find a model (function) that best describes the dependent with the independent variables, that is, to estimate parameters of the model,
2. To predict values of the dependent variable based on new measurements of the independent variables,
3. To analyze the importance of particular independent variables, thus, to analyze whether all or just some independent variables are important in the model.

Let the symbols y_i and x_i denote the measurements of weight and feed intake for animal I respectively. For n animals in this example the measurements are:

$Y=b_0+ b_1 (x) + e$, Where Y=dependent variable

X=independent variable

b_0= constant

b_1= regression coefficient

e= random error

Here, b_0 is called a constant and the parameter b_1 is called the regression coefficient, because it explains the slope. The random error ε is included in the model because changes of the values of the dependent variable are usually not completely explained by changes of values of the independent variable, but there is also an unknown part of that change. Regression parameters of a population are usually unknown, and they are estimated from data collected on a sample of the population.

Calculating of regression Analysis or prediction of equation for dependent variable from the independent one EX: Estimate the regression analysis of milk yield (Y) and feed intake (X) of five cows.

X	76	75	70	66	78
Y	15	14.8	13.8	13	15.5

To predict the regression equation, estimate slope of the relationship and the formula to estimate slope (b_1) is as follows: $b = \frac{\sum XiYi - \frac{(\sum Xi)(\sum Yi)}{n}}{\sum Xi_i^2 - \frac{(\sum Xi)^2}{n}}$

Similar to r, to calculate regression equation, we have to calculate the sum X, Y, XY and sum square of X and Y

Ser. No.	Y	X	Y^2	X^2	XY
1	15	76	225	5776	1140
2	14.8	75	219.04	5625	1110
3	13.8	70	190.44	4900	966
4	13	66	169	4356	858
5	15.5	78	240.25	6084	1209
$\sum$	72.1	365	1043.73	26741	5283

$$b = \frac{\sum XiYi - \frac{(\sum Xi)(\sum Yi)}{n}}{\sum Xi_i^2 - \frac{(\sum Xi)^2}{n}} = \frac{\sum 5283 - \frac{(365)(72.1)}{5}}{\sum 26741 - \frac{(365)^2}{5}} = \frac{5283 - 5263.3}{26741 - 26645} = 0.21$$

So, the value of b_1 = 0.21

Interpretation: when the amount of feed intake is increase by one unit (kg), the milk yield of a cow is increased by 0.21kg.

Calculate the intercept (b_0): After calculating the slope, the intercept is calculated from the following relationship: $\mathbf{b_0} = \bar{Y} - \mathbf{b_1}\bar{X}$

$\bar{Y} = \sum Yi/n$ =72.1/5 = 14.42

$\bar{X} = \sum Xi/n$ = 365/5 = 73

$b_0 = 14.42 - 0.21\ (73)$

$b_0 = 14.42 - 15.33$

$\mathbf{b_0} = -\mathbf{0.91}$

Accordingly the equation that describes the relationship between milk yield of cows and feed intake is: **Y= – 0.91 + 0.21(X)**

When X=85 and b_0=0.91; what was the value of Y?

Differences between correlation and regression

- **Linear correlation** is a statistic that measures the strength of two independent (X) and dependent (Y) variable association while;
- **Linear regression** is a prediction equation that estimates the value of Y for any given X.