भाकृअनुप
ICAR

हर कदम, हर डगर
किसानों का हमसफर
भारतीय कृषि अनुसंधान परिषद

Agrisearch with a human touch

Course on

Statistics: Experimental Designs and Analysis

for M.Sc. Students of
Afghanistan National Agricultural Sciences and Technology University (ANASTU)

(April 13 – May 8, 2020)

Compiled and Edited By

Seema Jaggi
Rajender Parsad
Sukanta Dash
Arpan Bhowmik

भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
INDIAN AGRICULTURAL STATISTICS
RESEARCH INSTITUTE

Teaching Manual

ICAR- Indian Agricultural Statistics Research Institute
Library Avenue, New Delhi – 110 012 (India)
www.iasri.res.in

PREFACE

Applications of appropriate experimental designs and statistical techniques forms the backbone of any research endeavour in agriculture and allied sciences. In order to maintain and improve the quality of agricultural research, it is of paramount importance that sound and modern statistical methodologies are used in the collection and analysis of data and then in the interpretation of results. The use of efficient and cost effective designs and appropriate statistical techniques for analyzing the data are very crucial to obtain a meaningful interpretation of the investigation. In this endeavor, ICAR-Indian Agricultural Statistics Research Institute, New Delhi has established itself in the field of Agricultural Statistics in general and Design of Experiments in particular.

This teaching manual on **Statistics: Experimental Designs and Analysis** has been prepared for students of **Afghanistan National Agricultural Sciences and Technology (ANASTU), Afghanistan** under a course for their M.Sc. programme in collaboration with PG School IARI, New Delhi. Total 26 students with 9 from Plant Protection and 17 from Horticulture disciplines of ANASTU attended this course scheduled from April 13 to May 8, 2020. Due to the COVID-19 pandemic during this period, the course was taken through online mode. The manual contains the lecture notes on different topics covered during this course starting from the basic statistical methods, testing of hypothesis, efficient design of experiments and analytical techniques of experimental data to multivariate statistical techniques along with some other useful statistical tools like data diagnostics and transformation, probit analysis, logistic regression, non-parametric test etc.. Emphasis has been also given on interpretation and presentation of results. Notes on MS-Excel and R along with online tools in the field of design of experiments that have been used for practical exercises have also been included. We are sure that this manual will be very much useful for the students in their current and future research studies.

We take this opportunity to thank all the students who have attended the course through online mode with full devotion and energy. Although every editorial care has been taken in compiling the teaching manual from available lecture notes of different faculty of ICAR-IASRI, New Delhi, errors and omissions are likely to occur. We welcome the constructive suggestions on any modifications/ improvements in this manual. We are thankful to ANASTU and Professor Anupam Varma for having faith on us for organizing this course. Our acknowledgements to Dr. VK Baranwal, Professor (Plant Pathology) and Dr. TK Behera, Professor (Vegetable Science). We are also grateful to Director, ICAR-IARI, Dean PG School, IARI and Director, ICAR-IASRI, for their full support for undertaking this course. We are also thankful to one and all for their efforts and help in preparing this manual.

New Delhi **Course Instructors**
08 May, 2020

CONTENTS

DESCRIPTIVE STATISTICS

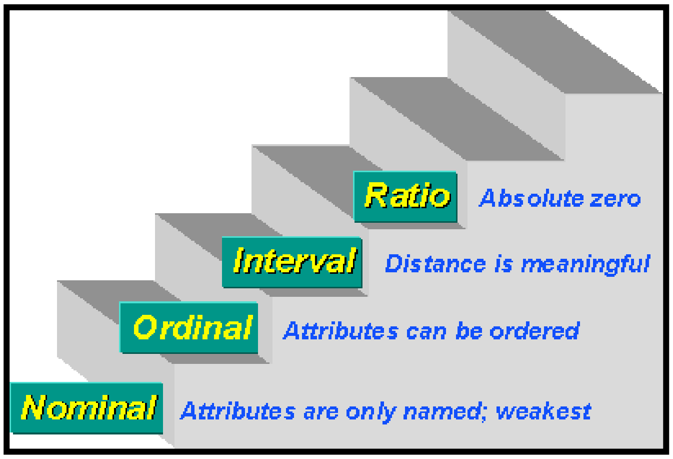
1. Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study, there may be lots of measures or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. There are two basic methods: numerical and graphical. Using the numerical approach one might compute statistics such as the mean and standard deviation. These statistics convey information about the average degree of shyness and the degree to which people differ in shyness. Graphical methods are better suited than numerical methods for identifying patterns in the data. Numerical approaches are more precise and objective. Since the numerical and graphical approaches compliment each other, it is wise to use both.

The raw data consist of measurements of some attribute on a collection of individuals. The measurement would have been made in one of the following scales *viz*., nominal, ordinal, interval or ratio scale.

2. Levels of Measurement

- **Nominal scale** refers to measurement at its weakest level when number or other symbols are used simply to classify an object, person or characteristic, *e.g.*, state of health (healthy, diseased).
- **Ordinal scale** is one wherein given a group of equivalence classes, the relation greater than holds for all pairs of classes so that a complete rank ordering of classes is possible, *e.g.*, socio-economic status.
- When a scale has all the characteristics of an ordinal scale, and when in addition, the distances between any two numbers on the scale are of known size, **interval scale** is achieved, e.*g*., temperature scales like centigrade or Fahrenheit.
- An interval scale with a true zero point as its origin forms a ratio scale. In a **ratio scale**, the ratio of any two scale points is independent of the unit of measurement, e.g., height of trees.



3. Types of Descriptive Statistics

- Graphs and Frequency Distribution
  These represent the data enabling the researcher to see what the distribution of scores look like.
- Measures of Central Tendency
  These measures are the indices that enable to determine the average score of a group of scores.
- Measures of Variability
  These measures are indices that enable to indicate how spread out a group of scores are.

4. Frequency Distribution

The frequency distribution is a summary of the frequency of individual values or ranges of values for a variable. Preparation of frequency distribution is an often-used technique in statistical works when summarizing large masses of raw data, which leads to information on the pattern of occurrence of predefined classes of events.

Ungrouped Data: The simplest distribution would list every value of a variable and the number of persons who had each value.

Grouped Data: A way to summarize data is to distribute it into **classes** or **categories** and to determine the number of individuals belonging to each class, called the **class frequency**. It is easier to see patterns in the data, but there is loss of information about individual scores.

A tabular arrangement of data by classes together with the corresponding class frequencies is called a **frequency distribution** or **frequency table**. Following is the raw data on some measurements and its frequency distribution:

86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	74	39	67
79	83	70	78	91
68	49	56	94	81

Table 1: Grouped frequency distribution

Class Interval	Frequency	Proportion	Cumulative Frequency
20-under 30	1	0.028	1
30-under 40	1	0.028	2
40-under 50	2	0.057	4
50-under 60	3	0.086	7
60-under 70	5	0.143	12
70-under 80	10	0.287	22
80-under 90	8	0.228	30
90-under 100	5	0.143	35

Following is a frequency distribution of Diameter at Breast Height (DBH) recorded to the nearest cm, of 80 teak trees in a sample plot.

Table 2: Frequency distribution of DBH of teak trees in a plot

DBH class (cm)	**Frequency (Number of trees)**
11 - 13	11
14 - 16	20
17 - 19	30
20 - 22	15
23 - 25	4
Total	80

5. Graphical Representation of Data

Frequency distributions are often graphically represented by a **histogram** or **frequency polygon**. A histogram consists of a set of rectangles having bases on a horizontal axis (the *x* axis) with centres at the class marks and lengths equal to the class interval sizes and areas proportional to class frequencies. If the class intervals all have equal size, the heights of the rectangles are proportional to the class frequencies and it is then customary to take the heights numerically equal to the class frequencies. If class intervals do not have equal size, these heights must be adjusted. A frequency polygon is a line graph of class frequency plotted against class mark. It can be obtained by connecting midpoints of the tops of the rectangles in the histogram.

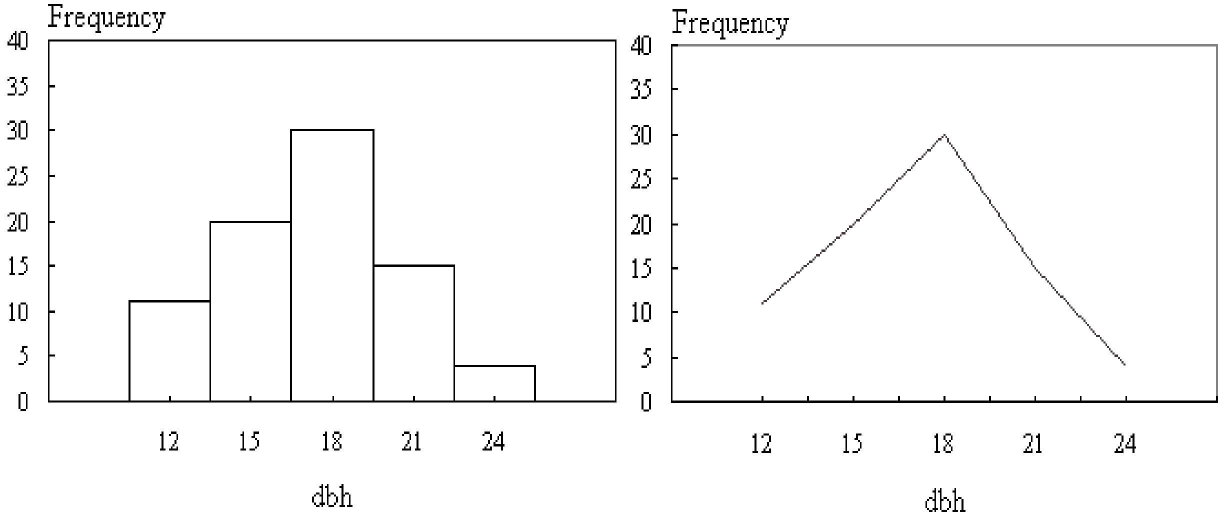


Fig. 1: Histogram and frequency curce showing the frequency distribution of DBH

The qualitative data is summarized in a frequency, relative frequency, or percent frequency distribution using **bar chart**. On the horizontal axis we specify the labels that are used for each of the classes. A frequency, relative frequency, or percent frequency scale is used for the vertical axis. Using a bar of fixed width drawn above each class label, the height can be extended appropriately. The bars are separated to emphasize the fact that each class is a separate category.

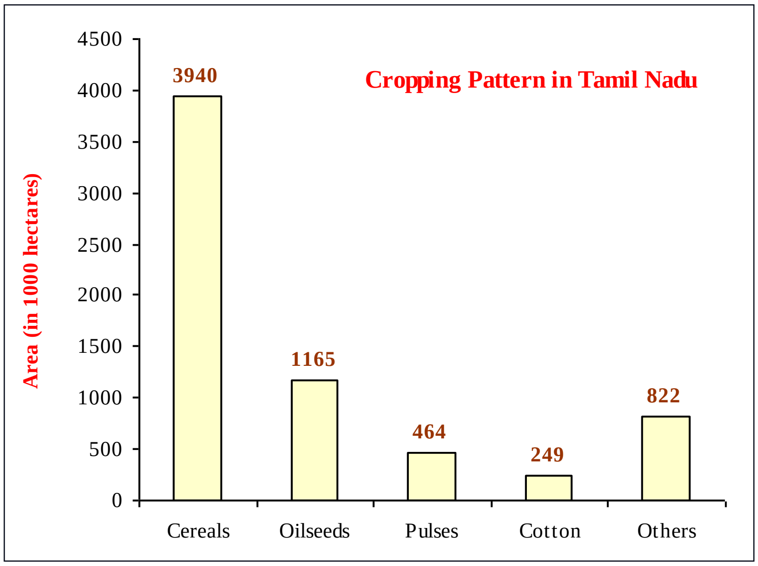


Fig. 2: Bar chart of cropping pattern

Pie chart is commonly used graphical device for presenting relative frequency distributions for qualitative data. Draw a circle; then use the **relative frequencies** to subdivide the circle into sectors that correspond to the relative frequency for each class. Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume .25(360) = 90 degrees of the circle. The above given cropping pattern is displayed in pie chart as follows:

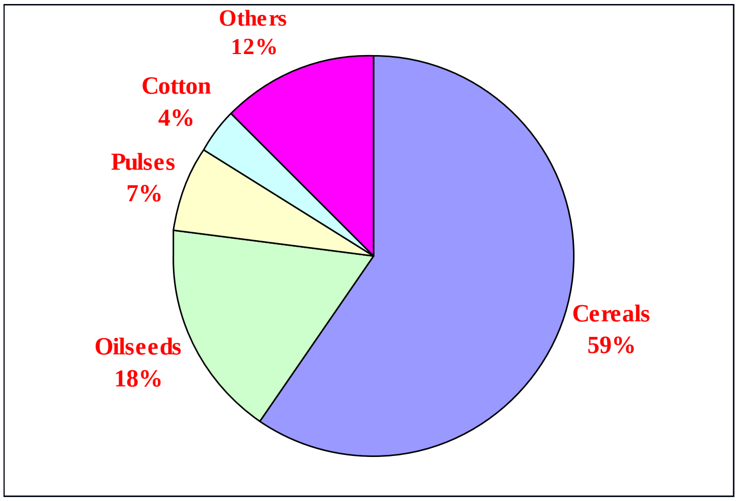


Fig. 3: Pie chart of cropping pattern

Having prepared a frequency distribution, a number of measures can be generated out of it, which leads to further condensation of the data. These are measures of location or central tendency, dispersion, skewness and kurtosis.

6. Measures of Central Tendency

The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean add up all the values and divide by the number of values. The arithmetic mean $(\bar{x})$ or the mean of a set of N numbers x_1, x_2, x_3,…, x_N is

$$\text{Mean} = \frac{x_1 + x_2 + ... + x_N}{N}$$

If the numbers x_1, x_2,…, x_k occur f_1, f_2,…,f_k times respectively i.e., occur with frequencies f_1, f_2, …, f_k, the arithmetic mean is

$$\text{Mean} = \frac{f_1x_1 + f_2x_2 + ... + f_kx_k}{f_1 + f_2 + ... + f_k}$$

Consider the data given in Table 2,

DBH class (cm)	**Frequency (f) (Number of trees)**	**x**	**xf**
11 - 13	11	12	132
14 - 16	20	15	300
17 - 19	30	18	540
20 - 22	15	21	315
23 - 25	4	24	96
Total	80	80	1383

$$\text{Mean} = \frac{1383}{80} = 17.29\text{cm}.$$

The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, let 8 scores be ordered as 15, 15, 15, 20, 20, 21, 25, 36. Score number 4 and 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, then average of two would determine the median.

For grouped data, the median is obtained using following:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - (\sum f)_1}{f_m} \right) c,$$

where L is lower class limit of the median class (i.e., the class containing the median), $(\sum f)_1$ is sum of frequencies of all classes lower than the median class, f_m is the frequency of median class and c is the class interval.

Geometrically, the median is the value of x (abscissa) corresponding to that vertical line which divides a histogram into two parts having equal areas.

For Table 2,

DBH class (cm)	**x**	**Frequency (f) (Number of trees)**	**Cumulative Frequency**
11 - 13	12	11	11
14 - 16	15	20	31
17 - 19	18	30	61
20 - 22	21	15	76
23 - 25	24	4	80
Total	80	80	

N / 2 = 40 which falls in the class 17-19 and is thus the median class.

$$\text{Median} = 16.5 + \left(\frac{\frac{80}{2} - 31}{30} \right) 3 = 17.4 \text{ cm.}$$

The **Mode** is the most frequently occurring value in the set of scores. To determine the mode, order the scores and then count each one. The most frequently occurring value is the mode. In the example 15, 15, 15, 20, 20, 21, 25, 36, the value 15 occurs three times and is the mode. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently. The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 has two modes 4 and 7 and is called bimodal.

In case of grouped data, the mode will be the value (or values) of x corresponding to the maximum point (or points) on the curve. From a frequency distribution or histogram, the mode can be obtained from the formula,

$$\text{Mode} = L + \left(\frac{f_2}{f_1 + f_2} \right) c \,,$$

where L is the lower class limit of modal class (the class containing the mode), f_1 is the frequency of the class previous to the modal class, f_2 is frequency of the class just after the modal class and c is the size of modal class.

From Table 2, the maximum frequency is 30 and hence the modal class is 17-19.

$$\text{Mode} = 16.5 + \left(\frac{15}{15 + 20} \right) 3 = 17.79 \text{ cm.}$$

Notice that for the same set of scores, we may get different values for the mean, median and mode. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other. With three different measures of central tendency, how to know which one to use? The answer depends a lot on the data and what is to be communicated.

While the mean is the most frequently used measure of central tendency, it does suffer from one major drawback. Unlike other measures of central tendency, the mean can be influenced profoundly by one extreme data point (referred to as an "outlier"). The median and mode clearly do not suffer from this problem. There are certainly occasions where the mode or median might be appropriate. For qualitative and categorical data, the mode makes sense, but the mean and median do not. For example, when we are interested in knowing the typical soil type in a locality or the typical cropping pattern in a region we can use mode. On the other hand, if the data is quantitative one, we can use any one of the averages.

If the data is quantitative, then one has to consider the nature of the frequency distribution. When the frequency distribution is skewed (not symmetrical) the median or mode will be proper average. In case of raw data in which extreme values, either small or large, are present, the median or mode is the proper average. In case of a symmetrical distribution either mean or median or mode can be used. However, as seen already, the mean is preferred over the other two. The mean, median, and mode can be related (approximately) to the histogram: the mode is the highest bump, the median is where half the area is to the right and half is to the left, and the mean is where the histogram would balance.

The **Harmonic mean** H of the positive real numbers x_1, x_2, ..., x_n is defined to be

$$H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}}$$

Equivalently, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. If a set of weights $w_1,...,w_n$ is associated to the dataset $x_1,...,x_n$, the weighted harmonic mean is defined by

$$H=\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i/x_i}$$

The **geometric mean** of a data set x_1, ..., x_n is given by

$$G=(x_1 x_2 ... x_n)^{1/n}$$

When dealing with rates, speed and prices, harmonic mean may be used. If interested in relative change, as in the case of bacterial growth, cell division etc., geometric mean is the most appropriate average.

7. Measures of Dispersion

Averages are representatives of a frequency distribution but they fail to give a complete picture of the distribution. They do not tell anything about the scatterness of observations within the distribution.

Suppose that we have the distribution of the yields (kg per plot) of two paddy varieties from 5 plots each. The distribution may be as follows:

Variety I	45	42	42	41	40
Variety II	54	48	42	33	30

It can be seen that the mean yield for both varieties is 42 kg. But we can not say that the performance of the two varieties are same. There is greater uniformity of yields in the first variety whereas there is more variability in the yields of the second variety. The first variety may be preferred since it is more consistent in yield performance. From the above example, it is obvious that a measure of central tendency alone is not sufficient to describe a frequency distribution. In addition to it, a measure of **scatterness** of observations should be there. The scatterness or variation of observations from their average is called the **dispersion**. There are different measures of dispersion like the range, the quartile deviation, the mean deviation and the standard deviation.

The **Range** is simply the highest value minus the lowest value. The **Standard Deviation** (S.D) is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range. The Standard Deviation shows the relation that set of scores has to the mean of the sample. The standard deviation is the square root of the sum of the squared deviations from the mean divided by the number of scores.

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

If x_1, x_2,…,x_k occur with frequencies f_1, f_2,…,f_k respectively, the standard deviation can be computed as

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{N}}\ , \ = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{N} - \left(\frac{\sum_{i=1}^{k} f_i x_i}{N}\right)^2}\ , \ N = \sum_{i=1}^{k} f_i$$

Consider the data given in Table 2.

DBH class (cm)	**Frequency (f) (Number of trees)**	**x**	**fx**	**fx^2**
11 - 13	11	12	132	1584
14 - 16	20	15	300	4500
17 - 19	30	18	540	9720
20 - 22	15	21	315	6615
23 - 25	4	24	96	2304
Total	80	80	1383	24723

$$\text{Standard Deviation} = \sqrt{\frac{24723}{80} - \left(\frac{1383}{80}\right)^2} = 3.19 \text{ cm.}$$

The **variance** of a set of data is defined as the square of the standard deviation. **Mean deviation** is the mean of the deviations of individual values from their average. The average may be either mean or median. For raw data the mean deviation from the median is the least.

Measures of Relative Dispersion

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they are expressed in different units of measurement, we can not use the standard deviations as such for comparing their variability. We have to use the relative measures of dispersion in such situations.

There are relative dispersions in relation to range, the quartile deviation, the mean deviation, and the standard deviation. Of these, the coefficient of variation which is related to the standard deviation is important. The ratio of standard deviation (S.D) to mean expressed in percentage is called **coefficient of variation**,

C.V. = (S.D. / Mean) x 100

The C.V. is a unit-free measure. It is always expressed as percentage. The C.V. will be small if the variation is small. Of the two groups, the one with less C.V. is said to be more consistent.

The coefficient of variation is unreliable if the mean is near zero. Also it is unstable if the measurement scale used is not ratio scale. The C.V. is informative if it is given along with the mean and standard deviation. Otherwise, it may be misleading.

Suppose that the variation in height of seedlings and that of older trees of a species are to be compared. Let the mean height of seedlings be 50 cm and standard deviation of height of seedlings be 10 cm. Further let the mean height of trees be 500 cm with standard deviation of height of seedlings as 100 cm. By the absolute value of the standard deviation, one may tend to judge that variation is more in the case of trees but the relative variation, as indicated by the coefficient of variation (20%), is the same in both the sets.

Consider the measurements on yield and plant height of a paddy variety. The mean and standard deviation for yield are 50 kg and 10 kg respectively. The mean and standard deviation for plant height are 55 cm and 5 cm, respectively.

Here the measurements for yield and plant height are in different units. Hence, the variability can be compared only by using coefficient of variation. For yield,

C.V. = (10 / 50) x 100 = 20 %

For plant height,

C.V. = (5 / 55) x 100 = 9.1 %

The yield is subject to more variation than the plant height.

8. Shape of the Distribution

An important aspect of the "description" of a variable is the shape of its distribution, which tells the frequency of values from different ranges of the variable. A researcher is interested in how well the distribution can be approximated by the normal distribution Simple descriptive statistics can provide some information relevant to this issue. For example, if the **skewness** (which measures the deviation of the distribution from symmetry) is clearly different from 0, then that distribution is **asymmetrical**, while normal distributions are perfectly symmetrical. If the

kurtosis (which measures "peakedness" of the distribution) is clearly different from 0, then the distribution is either flatter or more peaked than normal; the kurtosis of the normal distribution is 0.

Skewness is the degree of asymmetry, or departure from symmetry, of a distribution. If the frequency curve (smoothed frequency polygon) of a distribution has a longer ‘tail’ to the right of the central maximum than to the left, the distribution is said to be skewed to the right or to have positive skewness. If the reverse is true, it is said to be skewed to the left or to have negative skewness. An important measure of skewness expressed in dimensionless form is given by

$$\text{Cefficient of skewness} = \beta_1 = \frac{\mu_3^2}{\mu_2^3},$$

where μ_2 and μ_3 are the second and third central moments defined using the formula,

$$\mu_r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^r}{N}.$$

For grouped data, the above moments are given by

$$\mu_r = \frac{\sum_{i=1}^{N} f_i (x_i - \bar{x})^r}{N}.$$

For a symmetrical distribution, $\beta_1= 0$. Skewness is positive or negative depending upon whether β_1 is positive or negative.

Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution. A distribution having a relatively high peak is called leptokurtic, while the curve which is flat-topped is called platykurtic. A bell shaped curve which is not very peaked or very flat-topped is called mesokurtic. The measure of kurtosis, expressed in dimensionless form, is given by

$$\text{Cefficient of kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2},$$

where μ_4 and μ_2 can be obtained from the formula as given above. The distribution is called normal if $\beta_2 = 3$. When β_2 is more than 3, the distribution is said to be leptokurtic. If β_2 is less than 3, the distribution is said to be platykurtic.

PROBABILITY DISTRIBUTIONS

The concept of probability plays an important role in all problems of science and every day life that involves an element of uncertainty. **Probabilities** are defined as relative frequencies, and to be more exact as limits of relative frequencies. The relative frequency is nothing but the proportion of time an event takes place in the long run. When an experiment is conducted, such as tossing coins, rolling a die, sampling for estimating the proportion of defective units, several outcomes or events occur with certain probabilities. These events or outcomes may be regarded as a variable which takes different values and each value is associated with a probability. The values of this variable depend on chance or probability. Such a variable is called a **random variable**. Random variables which take a finite number of values or to be more specific those which do not take all values in any particular range are called **discrete** random variables. For example, when 20 coins are tossed, the number of heads obtained is a discrete random variable and it takes values 0,1,...,20. These are finite number of values and in this range, the variable does not take values such as 2.8, 5.7 or any number other than a whole number. In contrast to discrete variable, a variable is **continuous** if it can assume all values of a continuous scale. Measurements of time, length and temperature are on a continuous scale and these may be regarded as examples of continuous variables. A basic difference between these two types of variables is that for a discrete variable, the probability of it taking any particular value is defined. For continuous variable, the probability is defined only for an interval or range. The frequency distribution of a discrete random variable is graphically represented as a histogram, and the areas of the rectangles are proportional to the class frequencies. In continuous variable, the frequency distribution is represented as a smooth curve.

Frequency distributions are broadly classified under following two heads:

Observed frequency distributions and
Theoretical or Expected frequency distributions

Observed frequency distributions are based on observations and experimentation. As distinguished from this type of distribution which is based on actual observation, it is possible to deduce mathematically what the frequency distributions of certain populations should be. Such distributions as are expected from on the basis of previous experience or theoretical considerations are known as **theoretical distributions** or **probability distributions**.

Probability distributions consist of mutually exclusive and exhaustive compilation of all random events that can occur for a particular process and the probability of each event's occurring. It is a mathematical model that represents the distributions of the universe obtained either from a theoretical population or from the actual world, the distribution shows the results that are obtained if many probability samples are taken and the statistics is computed for each sample. A table listing all possible values that a random variable can take on together with the associated probabilities is called a probability distribution.

The probability distribution of X, where X is the number of spots showing when a six-sided symmetric die is rolled is given below:

X	1	2	3	4	5	6
f(X)	1/6	1/6	1/6	1/6	1/6	1/6

The probability distribution is the outcome of the different probabilities taken by the function of the random variable X.

Knowledge of the expected behaviour of a phenomenon or the expected frequency distribution is of great help in a large number of problems in practical life. They serve as benchmarks against which to compare observed distributions and act as substitute for actual distributions when the latter are costly to obtain or cannot be obtained at all.

We now introduce a few discrete and continuous probability distributions that have proved particularly useful as models for real-life phenomena. In every case the distribution will be specified by presenting the probability function of the random variable.

DISCRETE PROBABILITY DISTRIBUTIONS

Uniform Distribution

A uniform distribution is one for which the probability of occurrence is the same for all values of X. It is sometimes called a rectangular distribution. For example, if a fair die is thrown, the probability of obtaining any one of the six possible outcomes is 1/6. Since all outcomes are equally probable, the distribution is uniform.

Definition: If the random variable X assumes the values $x_1,x_2,...,x_k$ with equal probabilities, then the discrete uniform distribution is given by

$$P(X = x_i) = \frac{1}{k} \quad \text{for } i = 1,2,...,k$$

Example: Suppose that a plant is selected at random from a plot of 10 plants to record the height. Each plant has the same probability 1/10 of being selected. If we assume that the plants have been numbered in some way from 1 to 10, the distribution is uniform with f(x; 10) = 1/10 for x = 1,...,10.

Binomial Distribution

Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives i.e. success or failure. More precisely, the binomial distribution refers to a sequence of events which posses the following properties:

1. An experiment is performed under same conditions for a fixed number of trials say, n.
2. In each trial, there are only two possible outcomes of the experiment ‘success’ or ‘failure’.
3. The probability of a success denoted by p remains constant from trial to trial.
4. The trials are independent i.e. the outcomes of any trial or sequence of trials do not affect the outcomes of subsequent trials.

Consider a sequence of n independent trials. The interest is in the probability of x successes from n trials, a binomial distribution is obtained where x takes the values from 0,1,…,n.

Definition: A random variable X is said to follow a binomial distribution with parameters n and p if its probability function is given by

$$P[X = x] = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,...,n, \; 0 < p < 1.$$

The probability of success are the successive terms of the binomial expansion $(q+p)^n$. The probable frequencies of the various outcomes in N sets of n trials are $N(q+p)^n$. The frequencies obtained by this expression are known as expected or theoretical frequencies. The frequencies actually obtained by making experiments are called observed frequencies. Generally, there is some difference between the observed and expected frequencies but the difference becomes smaller and smaller as N increases.

The various constants of the binomial distribution are as follows:

Mean = np
Variance = npq (mean > variance)
First moment $\mu_1 = 0$
Second moment $\mu_2 = npq$
Third moment $\mu_3 = npq(q-p)$
Fourth moment $\mu_4 = 3n^2p^2q^2 + npq(1-6pq)$

$$\beta_1 = \frac{(q-p)^2}{npq} \; , \; \gamma_1 = \frac{q - p}{\sqrt{npq}}$$

$$\beta_2 = 3 + \frac{1 - 6pq}{npq} \; , \; \gamma_2 = \frac{1 - 6pq}{npq}$$

Properties of the binomial distribution

1. The shape and location of the distribution changes as p changes for a given n or as n changes for a given p. As p increases for a fixed n, the binomial distribution shifts to the right.
2. The mode of the binomial distribution is equal to the value of x which has the largest probability. The mean and mode are equal if np is an integer.
3. As n increase for a fixed p, the binomial distribution moves to right, flattens and spreads out. When p and q are equal, the distribution is symmetrical, for p and q may be interchanged without altering the value of any term, and consequently terms equidistant from the two ends of the series are equal. If p and q are unequal, the distribution is skewed. If p < 1/2, the distribution is positively skewed and when p > 1/2, the distribution is negatively skewed.
4. If n is large and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable given by

$$Z = \frac{X - np}{\sqrt{npq}} .$$

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, an experimenter wants to know the probability of obtaining diseased trees in a ra 4ndom sample of 10 trees if 10 percent of the trees are diseased. The answer can be obtained from the binomial probability distribution. The binomial distribution can be used to know the distribution of the number of seeds germinated out of a lot of seeds sown.

Illustration: The incidence of disease in a forest is such that 20% of the trees in the forest have the chance of being infected. What is the probability that out of six trees selected, 4 or more will have the symptoms of the disease?

Solution: The probability of a tree having being infected is

$$p = \frac{20}{100} = \frac{1}{5}$$

and the probability of not being infected = $1- \frac{1}{5} = \frac{4}{5}$

Hence the probability of 4 or more trees being infected out of 6 will be

$$P(X \geq 4) = \binom{6}{4}\left(\frac{1}{5}\right)^4\left(\frac{4}{5}\right)^{\mathbf{2}} + \binom{6}{5}\left(\frac{1}{5}\right)^{\mathbf{5}}\left(\frac{4}{5}\right)^1 + \binom{6}{6}\left(\frac{1}{5}\right)^6\left(\frac{4}{5}\right)^0$$

$$= \frac{53}{3125}.$$

Fitting a binomial distribution: When a binomial distribution is to be fitted to the observed data, the following procedure is adopted:

1. Evaluate mean of the given distribution and then determine the values of p and q. If one of these values is known, the other can be found out.
2. Expand the binomial $(q+p)^n$. The power n is equal to one less than the number of terms in the expanded binomial.
3. Multiply each term of the expanded binomial by N (the total frequency) in order to obtain the expected frequency in each category.

Exercise: The following data shows the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to this data.

X:	0	1	2	3	4	5	6	7	8	9	10
f:	6	20	28	12	8	6	0	0	0	0	0

Step 1: Calculate $\overline{X} = \frac{\sum fX}{\sum f}$

Step 2: Find p and q using mean = np.

Step 3: Expand the binomial $80(q+p)^{10}$ and find expected frequencies.

The generalization of the binomial distribution is the **multinomial distribution**. Whereas in case of binomial distribution, there are only two possible outcomes on each experimental trial, in the multinomial distribution there are more than two possible outcomes on each trial. The assumptions underlying the multinomial distribution are analogous to the binomial distribution. These are:

1. An experiment is performed under the same conditions for a fixed number of trials, say, n.
2. There are k outcomes of the experiment which may be referred to $e_1, e_2, e_3,...,e_k$. Thus the sample space of possible outcomes on each trial shall be:

   $S = \{e_1, e_2, e_3,...,e_k\}$
3. The respective probabilities of the various outcomes i.e., $e_1, e_2, e_3,...,e_k$ denoted by $p_1,p_2, p_3,...,p_k$ respectively remain constant from trial to trial. $(p_1+p_2+p_3+...+p_k=1)$
4. The trials are independent.

<u>Poisson Distribution</u>

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. This distribution is the limiting form of the binomial distribution as n becomes infinitely large and p approaches to zero in such a way that np = λ remains constant. A Poisson distribution may be expected in cases where the change of any individual event being a success is small. The distribution is used to describe the behaviour of rare events.

Definition: A random variable X is said to follow a Poisson distribution with parameter λ if the probability function is given by

$$P[X = x] = \frac{e^{-\lambda}\,\lambda^{x}}{x!},\ x = 0,1,... \qquad \text{where } e = 2.7183$$

The various constants of the Poisson distribution are

Mean = λ

Variance = λ (mean = variance)

First moment $\mu_1 = 0$

Second moment $\mu_2 = \lambda$

Third moment $\mu_3 = \lambda$

Fourth moment $\mu_4 = \lambda + 3\lambda^2$

$\beta_1 = \frac{1}{\lambda}$, $\gamma_1 = \frac{1}{\sqrt{\lambda}}$

$\beta_2 = 3 + \frac{1}{\lambda}$, $\gamma_2 = \frac{1}{\lambda}$

Properties of the Poisson distribution

1. As λ increases, the distribution shifts to the right, i.e. the distribution is always a skewed distribution.
2. Mode: When λ is not an integer then unique mode i.e. m = $[\lambda]$. When λ is an integer then bimodal i.e. m = λ and m = λ-1.
3. Poisson distribution tends to normal distribution as λ becomes large.

In general, the Poisson distribution explains the behaviour of discrete variates where the probability of occurrence of the event is small and total number of possible cases is sufficiently large. For example, it is used in quality control statistics to count the number of defects of an item, or in biology to count the number of bacteria, or in physics to count the number of particles emitted from a radioactive substance, or in insurance problems to count the number of casualties etc. The Poisson distribution is also used in problems dealing with the inspection of manufactured products with the probability that any piece is defective is very small and the lots are very large. Also used to know the probability of mutations in a DNA segment.

Note also that the only variable needed to generate these distributions is λ, the average occurrence/interval. Moreover, in biology situations often occur where knowing the probability of no events P(0) in an interval is useful. When x = 0, equation simplifies to P(0) =$e^{-\lambda}$. For example, we might want to know the fraction of uninfected cells for a known average (λ) multiplicity of virus infection (MOI). Other times, we need to know the average mutation rate/base pair, but our sequencing determines nearly all wild type sequence, P(0). In each case, if we can determine either λ or P(0), we can solve for the other.

Fitting a Poisson distribution: The process of fitting a Poisson distribution involves obtaining the value of λ, i.e., the average occurrence, and to calculate the frequency of 0 success. The other frequencies can be very easily calculated as follows:

$$N(P_0) = Ne^{-\lambda}$$

$$N(P_1) = N(P_0) \times \frac{\lambda}{1}$$

$$N(P_2) = N(P_1) \times \frac{\lambda}{2}$$

$$N(P_3) = N(P_2) \times \frac{\lambda}{3}\text{, etc.}$$

A ‘goodness-of fit’ test will confirm whether or not the fit is close enough to justify the belief that the distribution is of the Poisson type.

Exercise: The following mutated DNA segments were observed in 325 individuals:

Mutated DNA segments	0	1	2	3	4
Number of individuals	211	90	19	5	0

Fit a Poisson distribution to the data.
Step 1: Calculate the mean
Step 2: Find the different terms $N(P_0)$, $N(P_1)$,... i.e. the expected frequencies.

Negative Binomial Distribution

The negative binomial distribution is very much similar to the binomial probability model. It is applicable when the following conditions hold good:

1. An experiment is performed under the same conditions till a fixed number of successes, say c, are achieved.
2. In each trial, there are only two possible outcomes of the experiment ‘success’ or ‘failure’
3. The probability of a success denoted by p remains constant from trial to trial.
4. The trials are independent i.e. the outcome of any trial or sequence of trials do not affect the outcomes of subsequent trials.

The only difference between the binomial model and the negative binomial model is about the first condition.

Consider a sequence of Bernoulli trials with p as the probability of success. In the sequence, success and failure will occur randomly and in each trial the probability of success will be p. Let us investigate how much time will be taken to reach the rth success. Here r is fixed, let the number of failures preceding the rth success be x (=0,1,...). The total number of trials to be performed to reach the rth success will be x+r. Then the probability that rth success occurs at (x+r)th trial is

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r q^x \;; \quad x = 0,1,2,....$$

Illustration: Suppose that 30% of the items taken from the end of a production line are defective. If the items taken from the line are checked until 6 defective items are found, what is the probability that 12 items are examined?

Solution: Suppose the occurrence of a defective item is a success. Then the probability that there will be 6 failures preceding the 6th success will be given by:

$$\binom{6+6-1}{6-1}\ (.30)^6\ (.70)^6 = 0.0396.$$

If r = 1, i.e. the first success, then $P[X = x] = pq^x$, x=0,1,2,... which is the probability distribution of X, the number of failures preceding the first success. This distribution is called as **Geometric distribution**.

Hypergeometric Distribution

The hypergeometric distribution occupies a place of great significance in statistical theory. It applies to sampling without replacement from a finite population whose elements can be classified into two categories - one which possess a certain characteristic and another which does not possess that characteristic. The categories could be male-female, employed-unemployed etc.

When n random selections are made without replacement from the population, each subsequent draw is dependent and the probability of success changes on each draw. The following conditions characterise the hypergeometric distribution:

1. The result of each draw can be classified into one of the two categories.
2. The probability of a success changes on each draw.
3. Successive draws are dependent.
4. The drawing is repeated a fixed number of times.

Definition: The probability of r successes in a random sample of n elements drawn without replacement is;

$$P(r) = \frac{\binom{N - X}{n - r}\binom{X}{r}}{\binom{N}{n}} \quad \text{for } r=0,1,2...,[n,X]$$

The symbol [n, X] means the smaller of n or X.

This distribution may be used to estimate the number of wild animals in forests or to estimate the number of fish in a lake.

The hypergeometric distribution bears a very interesting relationship to the binomial distribution. When N increases without limit, the hypergeometric distribution approaches the binomial distribution. Hence, the binomial probabilities may be used as approximation to hypergeometric probabilities when n/N is small.

CONTINUOUS PROBABILITY DISTRIBUTION

Normal Distribution

The normal distribution is “probably” the most important distribution in Statistics. It is a probability distribution of a continuous random variable and is often used to model the distribution of discrete random variable as well as the distribution of other continuous random variables. The basic form of normal distribution is that of a bell, it has single mode and is symmetric about its central values. The flexibility of using normal distribution is due to the fact that the curve may be centered over any number on the real line and it may be made flat or peaked to correspond to the amount of dispersion in the values of random variable.

Many quantitative characteristics have distribution similar in form to the normal distribution's bell shape. For example height and weight of people, the IQ of people, height of trees, length of leaves etc. are typically the type of measurements that produces a random variable that can be successfully approximated by normal random variable. The values of random variables are produced by a measuring process and measurements tend to cluster symmetrically about a central value.

Definition: A normal distribution in a variate X with mean μ and variance σ^2 is a statistic distribution with probability function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

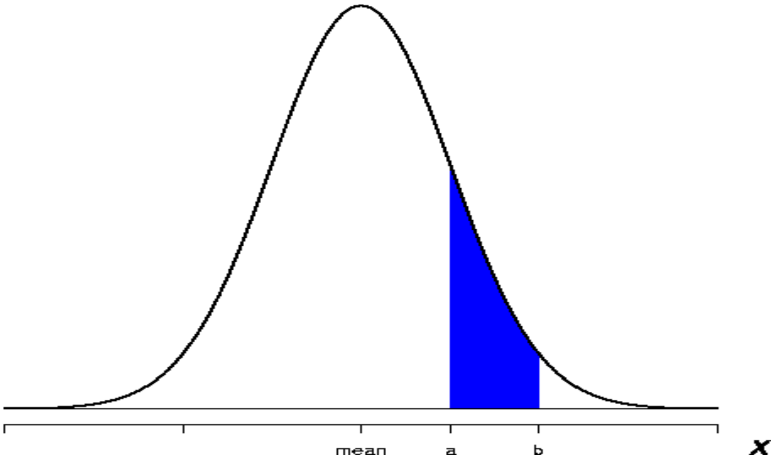
on the domain $x \in (-\infty, \infty)$. μ and σ^2 are parameters of the distribution.

If X is a normal random variable with mean μ and standard deviation σ, then $\frac{X-\mu}{\sigma}$ is a standard normal variate with zero mean and standard deviation 1. The probability density function of standard normal variable Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Area under the normal curve: For normal variable X,

P(a < X < b) = Area under f(x) from X = a to X = b



The probability that X is between a and b (b > a) can be determined by computing the probability that Z is between (a - μ) / σ and (b - μ) / σ. It is possible to determine the area in Fig. ii by using tables (for areas under normal curve) rather then by performing any mathematical computations.

Probability associated with a normal random variable X can be determined from Table 1 given at the end. As indicated in Fig. iii for any normal distribution, 68.27% of the Z values lie within one standard deviation of mean, 95.45% of values lie within 2 standard deviations of mean and 99.73% of values lie within three standard deviations of mean.

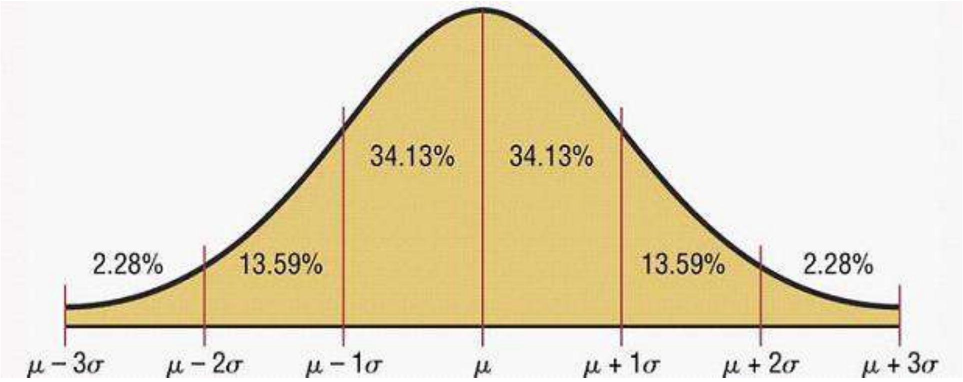


Fig. iii

The normal distribution is symmetric about its mean (zero in this case) and the total area under curve is 1 (half to the left of zero and half to right), Percentage points (right tail area) of normal distribution for various values of z are provided in Table 1 in the end.

Properties of normal distribution

1. The normal curve is symmetrical about the mean $x = \mu$.
2. The height of normal curve is at its maximum at the mean. Hence the mean and mode of normal distribution coincides. Also the number of observations below the mean in a normal distribution is equal to the number of observations above the mean. Hence mean and median of normal distribution coincides. Thus for normal distribution mean = median = mode.
3. The normal curve is unimodal at $x = \mu$.
4. The point of inflexion occurs at $\mu \pm \sigma$.
5. The first and third quartiles are equidistant from the median.
6. The area under normal curve is distributed as follows
   (a) $\mu \pm \sigma$ covers 68.27% of area
   (b) $\mu \pm 2\sigma$ covers 95.45% of area
   (c) $\mu \pm 3\sigma$ covers 99.73% of area

Importance of normal distribution

1. Of all the theoretical distributions, the normal distribution is the most important and is widely used in statistical theory and work. The most important use of normal distribution is in connection with generalization from a limited number of individuals observed on to individuals that have not been observed. It is because of this reason that the normal distribution is the core heart of sampling theory. The distribution of statistical measures such as mean or standard deviation tends to be normal when the sample size is large. Therefore, inferences are made about the nature of population from sample studies.
2. The normal distribution may be used to approximate many kinds of natural phenomenon such as length of leaves, length of bones in mammals, height of adult males, intelligence quotient or tree diameters. For example, in a large group of adult males belonging to the same race and living under same conditions, the distribution of heights closely resembles the normal distribution.
3. For certain variables the nature of the distribution is not known. For the study of such variables, it is easy to scale the variables in such a way as to produce a normal distribution.

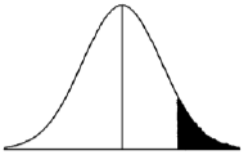


Table 1: Percentage points (right tail area) of normal distribution for various values of z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.2272	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2231	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1984	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0017	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

CORRELATION AND REGRESSION

1. Introduction

In order to study the relationship between two or more variables through correlation and or regression, it is important to visualize the relation between them graphically. Scatter Plot is the simplest way of the diagrammatic representation of a bivariate data. It gives the idea of the distribution of the data like well defined positive or negative linear relationships, non-linear relationships or no apparent relationship.

2. Scatter Plots

It is the simplest way of the diagrammatic representation of a bivariate data. It gives the idea of the distribution of the data like well defined positive or negative linear relationships, non-linear relationships or no apparent relationship. The chart can be created using the Graph menu. **To Obtain Scatterplots**: from the menus, choose: Graphs→ Scatter. SPSS 10.0 gives three types of scatter plots viz. *simple, overlay, matrix, or 3-D*. For getting the desired scatter plot click the icon and then Select Define → Select variables and options for the chart.

To Obtain Simple Scatterplots: From the menus, choose: Graphs → Legacy dialogs → Scatter/dot → Select the icon for Simple → Select Define → select a variable for the Y-axis and a variable for the X-axis. (Caution: These variables must be numeric, but should not be in date format). → if desired, select a variable and move it into *the Set Markers by* box. Each value of this variable is marked by a different symbol on the scatterplot. This variable may be numeric or string. → If desied, one can select a numeric or a string variable and move it into the *Label Cases* by box. You can label points on the plot with this variable.

- If selected, the value labels (or values if no labels are defined) of this variable are used as point labels.
- If we do not select a variable to label Cases by, case numbers can be used to label outliers and extremes.
- Select *Options* to specify the treatment of missing values in the data and control whether labels are to be displayed for points on the plot.
- Select *Titles* to define lines of text to be placed at the top or bottom of the plot.

To Obtain Overlay Scatterplots: This option is used to obtain plots for two or more variable pairs.

Select Graphs→ Legacy dialogs → Scatter→ Select the icon for Overlay→ Select Define → Select at least two pairs of variables, Select each variable separately. The first variable will appear in the Current Selections list box as Variable 1, and the second variable will appear as Variable 2. To deselect a variable, select it again in the source variable list. Once a pair of variables is selected, move the pair into the Y-X box (Caution: Variables should be numeric, but should not be in date format.). As in case of simple scatter plots, select a numeric or a string variable and move it into the Label Cases by box. Points on the plot are labeled with the selected variable.

To reverse the order of the Y and X variables within a selected pair, select Swap Pair. For the specification of the treatment of the missing values and case labels display and for titles, follow the steps as in simple scatter plot.

To Obtain a Scatterplot Matrix: This option plots all possible combinations of two or more numeric variables against one another. For obtaining a Scatterplot matrix, select the icon for Matrix → Select Define → Select at least two Matrix numeric Variables. Rest options are similar to the earlier ones.

To Obtain 3-D Scatterplots: This option plots three numeric variables in three dimensions. Select the icon for 3-D→ Select Define → Select one variable for the Y-axis, one for the X-axis and one for the Z-axis. These variables must be numeric, but should not be in date format.

3. Bivariate and Partial Correlation

A correlation coefficient measures the strength of a linear association between two quantitative variables. The most commonly used measure of linear correlation between two variables is called the *Pearson-product- moment correlation coefficient* or simply the *sample correlation coefficient* and is denoted by *r*. The values of the correlation coefficient is not expressed in units of the data, but range from –1 to +1. While scatterplot provide a picture of the relation, the value of the correlation is the same if you switch the *Y* (vertical) and *X* (horizontal) measures. The sample correlation coefficient *r* is estimated by the formula

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

For a sample of size *n*, the above expression can be written as

$$r = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where s_x and s_y are the sample standard deviations of the two variables. The formula can be simplified to

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}\,.$$

Test of significance of correlation coefficient

Case I: Let the population correlation coefficient of X and Y is denoted by ρ, then it is often of interest to test whether ρ is zero or different from zero, on the basis of observed correlation coefficient, r. Thus, if r is the sample correlation coefficient based on a sample of n observations, then the appropriate test procedure for testing the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ is:

1. Compute the quantity $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.
2. Compare the computed value of $|t|$, with the table value of *t*-distribution with $(n-2)$ degrees of freedom, and at a given level of significance, say 5 % .

3. If the computed value of $|t|$ exceeds the table value (as in (ii) above), then $H_0 : \rho = 0$ is rejected against the alternative $H_1 : \rho \neq 0$.

Case II: One may be interested in testing $H_0 : \rho = \rho_0$ against the alternative $H_1 : \rho \neq \rho_0$. This sample correlation coefficient based on *n* pairs of observations is based on the following quantity

$$\frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

which is a value of a random variable that follows approximately the normal distribution with mean $\frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$ and variance $1/(n-3)$. Thus the test procedure is to compute

$$Z = \frac{\sqrt{n-3}}{2}\left(\log_e\left(\frac{1+r}{1-r}\right) - \log_e\left(\frac{1+\rho_0}{1-\rho_0}\right)\right)$$

$$= \frac{\sqrt{n-3}}{2}\log_e\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0}\right]$$

and compare to the critical points of the standard normal distribution. For example, if the absolute value of Z, $|Z| > 1.96$, then the null hypothesis $H_0 : \rho = \rho_0$ against the alternative $H_1 : \rho \neq \rho_0$ is rejected at 5% level of significance. The alternative hypotheses $\rho < \rho_0$ or $\rho > \rho_0$ can also be tested using one tailed critical points.

Rank Correlation

In some cases, it is not possible to measure the data and only ranking is done. In such situations, the rank correlation is worked out which is nothing but the Pearson’s Product moment correlation coefficient and is defined as the correlation between ranks of individuals with respect to two characters. This is also known as *Spearman’s Rank correlation coefficient* and lies between –1 and +1. If d_i denotes the difference between the ranks of i^{th} individual and *n* denotes the number of individuals, then the Spearman’s Rank Correlation Coefficient is given by

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

If there is a tie in the ranks, then the ranks assigned is the average of the ranks assigned to these individuals had there been no tie. In case of ties, the rank correlation coefficient is given by

$$r = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2 + \sum (m^3 - m)/12\right)}{n(n^2-1)}$$, where *m* is the number of individuals having the same rank.

If in a group the data on more than two variables is collected and one is interested in obtaining the measure of linear association between all pairs of variables, then one can obtain the sample correlation coefficient for all possible pairs of variables. The probability of significance of each of these correlation coefficients can be obtained using any standard statistical software. However, if one scans the results for more than one pair of variables, the probabilities of significance are *pseudo probabilities* because they are designed to test one and only one correlation for significance and do not reflect the number of correlations tested. As a result some of the correlations may appear significant when they are not. The Bonferroni method may be used to adjust the stated significance levels. In this method, we divide the desired level of significance by *m* the number of correlation coefficients and if the probability is less than or equal to this ratio, then the correlation coefficient is significant at that level of significance. Alternatively, we multiply the probabilities of significance by *m* the number of correlation coefficients and if the probability is less than or equal to the desired level of significance, then the correlation coefficient is significant at that level of significance.

In the situations, when the number of observations or each pair of variables is not constant, one has to be cautious in scanning these *m* values to get a sense of the size of one correlation relative to another.

Partial Correlation

Sometimes the correlation between two variables Y and X_1 may be partly due to the correlation of a third variable, X_2 with both Y and X_1. The true correlation between Y and X_1 can only be observed once the effect of X_2 has been eliminated. We accomplish this by means of the *sample partial correlation coefficient*. Thus, partial correlation measures the linear association between two variables after the effects of one or more variables are removed. Partial correlation can reveal variables that enhance or suppress the relation between two particular variables. For example, if each Sunday for a year, one counts the number of ants in the kitchen at a beach cabin and the number of cars passing the house in a five-minute interval, the correlation may be close to 1. Are the cars bringing the ants? Does this sound silly? A third variable, temperature is ignored. When the weather is hot, the ants flourish and lots of people flock to the beach; when it is cool, the numbers of both the cars and ants diminish. If the linear effect of temperature is controlled, the relationship between ants and cars disappears.

The *partial correlation* of variables Y and X_1 after removing the effect of variable X_2 (or "controlling" for variable X_2) is estimated as follows:

- Regress variable Y on X_2.
- Regress variable X_1 on X_2.
- For each case, compute the residuals for each of the regression equations.
- Compute the usual Pearson correlation between the two sets of residuals.

The residuals represent variables Y and X_1 with the effect of variable X_2 removed. The partial correlation coefficient between Y and X_1 after eliminating the effect of variable X_2 is denoted by the symbol $r_{Y1.2}$. If we write the ordinary correlation coefficients for Y and X_1, Y and X_2, and X_1 and X_2 as r_{Y1}, r_{Y2}, and r_{12}, respectively, the sample partial correlation coefficient for Y and X_1 with X_2, held fixed is given by the following definition.

Partial Correlation Coefficient: The measure of linear relationship between the variable Y and X_1 after making allowance for their association with X_2, is estimated by the sample partial correlation coefficient $r_{y1.2}$, where

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2} r_{12}}{\sqrt{\left(1 - r_{Y2}^2\right)\left(1 - r_{12}^2\right)}}.$$

A similar definition applies to $r_{Y2.1}$ which measures the correlation between Y and X_2 after eliminating the linear effect of X_1.

The partial correlation coefficients obtained after removing the effect of one variable as discussed above are called partial correlation coefficients of order one. In some situations, however, we may have to obtain the partial correlation coefficients after eliminating the effects of two or more variables. The number of variables that are used for eliminating the effects is known as the order of the sample partial correlation coefficient.

Test of Significance of Partial Correlation Coefficient

To test $H_0 : \rho_{ij.12...} = 0$ against $H_1 : \rho_{ij.12...} \neq 0$ compute

$$t = \frac{r_{ij.12...}}{\sqrt{1 - r_{ij.12...}^2}} \sqrt{n - \theta - 2}$$

where θ is the order of the coefficient. This statistic follows t-distribution with $n - \theta - 2$ degrees of freedom. Reject H_0 if $|t| > t_{\alpha/2, n-\theta-2}$.

Steps to obtain correlation coefficient using MS-EXCEL: One can compute correlation coefficient by using Correl function in MS-EXCEL as CORREL(array1,array2), where Array1 is a cell range of values and Array2 is a second cell range of values.

One can also obtain bivariate correlations by using Tools → Data Analysis→ Correlation and then choosing the input and output range. For testing of significance or working out the exact probability level of significance one may use the following:

Probability level of significance can be obtained by TDIST(x,degrees_freedom,tails), x is the numeric value at which to evaluate the distribution, Degrees_freedom is an integer indicating the number of degrees of freedom and Tails specifies the number of distribution tails to return. If tails = 1, TDIST returns the one-tailed distribution. If tails = 2, TDIST returns the two-tailed distribution.

Alternatively, we can get the t-value of the Student's t-distribution as a function of the probability and the degrees of freedom by using TINV (probability, degrees_freedom). Here, probability is the probability associated with the two-tailed Student's t-distribution.

3. Regression

In many statistical studies, the goal is to establish a relationship, expressed via an equation, for predicting typical values of one variable given the value of another variable(s). In such situations, regression analysis can be of help to us. The term *regression* is derived from the original heredity studies made by Sir Francis Galton (1822-1911) in which he compared the heights of sons to the heights of fathers. Galton showed that the heights of the sons of tall fathers over successive generations regressed towards the mean height of the population. In other words, sons of usually tall fathers tend to be shorter than their fathers and sons of usually short fathers tend to be taller than their fathers. Now-a-days, the term regression is most of the prediction problems and does not necessarily imply a regression towards the population mean. In this section, we deal with the problem of estimating or predicting the value of a dependent variable given a set of independent variables. We begin with the case of single independent variable.

Simple Linear Regression

Let the variation in response variable (y) is explained by independent variable (x) called *regressor*. Simple regression of y on x or equation of a straight line as a statistical model, add a term for random error (ε) because the points do not fall on the line:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The *slope* (β_1) is the ratio between the vertical change and the horizontal change along the line. A test $\beta_1 = 0$ is same as that of a test that correlation coefficient (*r*) is zero as $r = \hat{\beta}_1 s_x / s_y$.

The *intercept* (β_0 or *constant* as it is often called) is where the line intercepts the vertical axis at *x = 0*.

To represent the errors (ε) in the model, draw a short vertical line from each point to the line. The lengths of these line segments between the line and the plot points are called *residuals* and are estimates for the true errors.

In the above equation, y is the *dependent* or *outcome* or *predicted* variable, the one you are trying to predict; x is the *independent* or *predictor* variable; and the intercept (β_0) and slope (β_1) are *coefficients*. If the model is a good descriptor of the relation between the variables, one can use the estimates of the coefficients to predict the value of the dependent variable for new cases.

Fitting of Simple Regression

Suppose *n* observations are made on y and x. Then, for each observation we have unobserved error term ε_i. We make the following assumptions regarding the ε_i's, which are random variables (i) errors are independent (ii) errors have zero mean and constant variance (σ^2). These assumptions can also be written as

$$E(\varepsilon_i) = 0,\ Var(\varepsilon_i) = \sigma^2 \quad \text{for all } i = 1, 2, \cdots, n.$$
$$Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \qquad \text{for all } i \neq i' = 1, 2, \ldots, n$$

Estimation of Parameters

The method of least squares for estimating the parameters β_0, β_1 as also σ^2, requires the minimization of the error sum of squares, *i.e.*, the sum of the squares of the vertical line segments, given by

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Differentiating S w.r.t. β_0 and β_1, and equating the derivatives to zero, we get a set of two equations as

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

These equations are called normal equations. The solution of these equations gives us the least squares estimates of β_0 and β_1 as b_0 and b_1

$$b_1 = Sxy / Sxx$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$

and $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \dfrac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}$. Further, let $s_{xy} = S_{xy} /(n-1)$ and $s_{xx} = S_{xx} /(n-1)$.

It should be noted that these estimates do minimize the error sum of squares, S. The fitted regression equation is thus

$$\hat{y} = b_0 + b_1 x$$

or $\hat{y} = \bar{y} + b_1 (x - \bar{x})$

The 'hat' over y indicates that if were substitute for x a value that is within the observed range of the predictor x, but has not necessarily been observed, then the regression equation gives us the predicated y for that given value of x. Note that if we set $x = \bar{x}$ in the fitted regression equation, then $y = \bar{y}$, meaning thereby that the point $\left(\bar{x}, \bar{y}\right)$ lies on the regression line.

Estimation of σ^2

In addition to estimating β_0 and β_1, an estimate of σ^2 is required to test hypotheses and construct interval estimates pertinent to the regression model. Ideally, we would like this estimate not to depend on the adequacy of the fitted model. This is only possible when there are several observations on y for at least one value of x, or when prior information concerning σ^2 is available. When this approach cannot be used, the estimate of σ^2 is obtained from the residual or error sum of squares.

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

A convenient computing formula for SSE may be found by substituting $\hat{y} = b_0 + b_1 x_1$ and simplifying, yielding

$$SSE = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b_1 S_{xy} = S_{yy} - b_1 S_{xy}$$

The residual sum of square has $\left(n-2\right)$ degrees of freedom, two degrees are associated with the estimates b_0 and b_1, involved in obtaining $\hat{y}_i$. Now the expected value of SSE is $E\left(SSE\right) = \left(n-2\right)\sigma^2$ so an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = MSE$$

The quantity MSE is called the error mean square or the residual mean square. The square root of s^2 is sometimes called the standard error of regression, and it has the same units as the response variable y. Because σ^2 depends on the residual sum of squares, any violation of the assumption on the model errors or any misspecification of the model form may seriously damage the usefulness of s^2 as an estimate of σ^2.

The above splitting of the total sum of squares due to $y's$ into two components can be formally put in an Analysis of variance table, as below:

Analysis of Variance: Simple Linear Regression

Source of Variation	**d.f.**	**S.S.**	**M.S.**
Regression	1	$b_1 S_{xy}$	MSR
Deviation form Regression (Residual)	$n-2$	SSE	$s^2 = SSE/\left(n-2\right) = MSE$
Total (corrected mean)	$n-1$	S_{yy}	

The total variation in y is partitioned in two parts as variations due to regression and deviation from regression. The test statistic is

$$F_0 = \frac{MSR}{MSE}$$

F_0 follows F_{1}, n_{-2} distribution of H_0: $\beta_1 = 0$ if $F_0 > F_{\alpha,1,} n_{-2}$.

We have seen above that $SSE = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b_1 S_{xy} = S_{yy} - b_1 S_{xy}$. It can further be simplified to

$$SSE = (n-1)(s_{yy} - b_1 s_{xy})$$

Now dividing both sides by $(n-1)s_{yy}$, we obtain

$$\frac{SSE}{(n-1)s_{yy}} = 1 - \frac{b_1 s_{xy}}{s_{yy}}$$

$$r^2 = 1 - \frac{SSE}{(n-1)s_{yy}}.$$

From the above, it can be concluded that r^2 measures the proportion of the total variation in the values of Y that can be accounted for or explained by the linear relationship with the values of X. Thus a correlation of $r = 0.6$ means that 0.36 or 36% of the total variation of the values of Y in our sample is accounted for by a linear relationship with the values of X. r^2 (Square of the correlation coefficient) is also known the coefficient of determination.

Remark 1: On the similar lines as above, the square of the sample partial correlation coefficient is called as sample coefficient of partial determination, which represents the ratio of the unexplained variation to the previously unexplained variation. That is $r^2_{Y1.2}$ gives us the proportion of the variation in the values of Y that was unexplained by a regression line involving only X_2 that can now be explained by including X_1 in the model along with X_2.

Precision of estimates

We derive the variances of b_1, b_0 and $\hat{y}_i$ for obtaining precision of estimates.

$$b_1 = S_{xy} / S_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$Var(b_1) = \frac{\sigma^2}{S_{xx}}$$

$$Var(b_0) = Var(\bar{y} - b_1\bar{x}) = Var(\bar{y}) + \bar{x}^2 \; v(b_1)$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2.\sigma^2}{S_{xx}} = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$$

Test of Significance of $\beta's$

We are often interested in testing hypothesis about model parameters. The tests are valid, if the assumption of normality of error terms is satisfied. One may be interested in testing the hypotheses $H_0 : \beta_0 = a$ against $H_1 : \beta_0 \neq a$. The appropriate test statistic for testing this is

$$t = \frac{|b_0 - a|}{SE(b_0)} = \frac{|b_0 - a|}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

Which follows, a t-distribution with $(n-2)$ d.f. If H_0 is true. Reject H_0 if $t > t_{\alpha/2, n-2}$.

We may be further interested in knowing, whether x is contributing significantly towards variability in y. This can be known by testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. We use the statistic

$$t = \frac{|b_1|}{SE(b_1)} = \frac{|b_1|}{\sqrt{\frac{s^2}{S_{xx}}}}$$

Which is distributed as t with $(n-2)$ d.f. If H_0 is true. Reject H_0 if $t > t_{\alpha/2, n_{-2}}$. Alternatively, the analysis of variance can also be used for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Variance of estimated mean and variance of prediction

The variance of $\hat{y}_i$ will be derived for the two situations where $\hat{y}_i$ is used as an estimate of the mean and where it is used as a prediction. Variance when $\hat{y}_i$ is used as the estimate of true mean of y at the specific value of x.

$$\hat{y}_i = b_0 + b_1 x_i = \bar{y} - b_1 (x_i - \bar{x})$$

$$V(\hat{y}_i) = \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{S_{xx}} = \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \sigma^2$$

The variance of the fitted value attains its minimum of σ^2 / n when the regression equation is evaluated at $x_i = \bar{x}$ and increases as the value of x moves away from $\bar{x}$.

When $\hat{y}_i$ is used as a predictor for some future observation, the variance for prediction must take into account the fact that the quantity being predicated is itself a random variable. Therefore, variance for prediction, $Var\left(y_{i(pred)}\right)$ is the variance of the difference between $\hat{y}_i$ and the future observation y_f

$$\begin{aligned} Var\left(y_{i(pred)}\right) &= V(\hat{y}_i - y_f) \\ &= V(\hat{y}_i) + \sigma^2 \\ &= \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \sigma^2 \end{aligned}$$

Note that the variance for prediction is the variance for estimation plus the variance of the quantity being predicted.

Example 3.1:

Data of illustration

Observation No.	y	x
1	78.5	7
2	74.3	1
3	104.3	11
4	87.6	11
5	95.9	7
6	109.2	11
7	102.7	3
8	72.5	1
9	93.1	2
10	15.9	21
11	83.8	1
12	113.3	11
13	119.4	10

Model to be fitted is $y = \beta_0 + \beta_1 x + \varepsilon$

Normal equations for estimation of parameters are

$$13b_0 + 97b_1 = 1250.5$$

$$97b_0 + 1139b_1 = 10132.0$$

These can also be written as

$$\begin{bmatrix} 13 & 97 \\ 97 & 1139 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1250.5 \\ 10132.0 \end{bmatrix}$$

Parameter estimates are

$$\underset{(5.437)}{b_0 = 81.792,} \qquad \underset{(0.581)}{b_0 = 1.930,} \qquad s^2 = 140.01$$

The figures in the parenthesis denote the SE of the estimated parameter.

Fitted model is $y = 81.792 + 1.930x$ $\qquad \left(r^2 = 0.501\right)$

Test of significance of $\beta' s$

(a) $H_0 : \beta_0 = 0,$ $\qquad H_1 : \beta_0 \neq 0,$

$$t = \frac{81.792}{5.437} = 15.043 \qquad \left(t_{.05'11} = 2.201\right)$$

(b) $H_0 : \beta_0 = 0,$ $\qquad H_1 : \beta_0 \neq 0,$

$$t = \frac{1.93}{.581} = 3.322$$

Estimated mean response at $x = 4$

$$\hat{y}_0 = 81.792 + 1.930(4) = 89.512$$

$$\hat{V}(\hat{y}_0) = 14.829 \qquad \text{Table value of } t_{\alpha/2,11} = 2.201$$

95% confidence interval for mean response y_0 is

$$89.512 - 2.201\sqrt{14.829} \le y_0 \le 89.512 + 2.201\sqrt{14.829}$$

$$89.512 \le y_0 \le 95.988$$

If $\hat{y}_0$ is used for prediction of future observation, then

$$\hat{V}(\hat{y}_0) = 140.101 + 14.829 = 154.930$$

95% confidence interval of prediction $\hat{y}_0$ is

$89.512 \pm 2.201\sqrt{154.930}$ i.e. 62.116 to 116.108.

Test for Linearity of Regression

For any given problem we assume the regression is linear and proceed with the estimation of parameters as discussed above. This assumption is made to avoid laborious calculations. A linear regression equation is always preferred over a nonlinear regression curve if the assumption of linearity can be justified. Therefore, the linearity of regression must be tested using the following test.

Let us select a random sample of n observations using k distinct values of x, says $x_{1,} x_2, \cdots, x_k$, such that the sample contains n_1 observed values of the random variable y_1 corresponding to x_1, n_2 observed values of y_2 corresponding to x_2, $\cdots$, n_k observed value of y_k corresponding to x_k, $n = \sum_{i=1}^{n} n_i$. We define

y_{ij} = jth value of the random variable y_i,

$y_{i.}$ = sum of the value of y_i in our sample.

Hence, if $n_4 = 3$ measurements of y are made corresponding of $x = x_4$, we could indicate these observations by y_{41}, y_{42}, and y_{43}. Then $y_{4.} = y_{41} + y_{42} + y_{43}$. Now the computed value

$$f = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)},$$

where $\chi_1^2 = \sum \frac{y_{i.}^2}{n_i} - \frac{\left(\sum y_{ij}\right)^2}{n} - b^2(n-1)s_x^2$

$$\chi_2^2 = \sum y_{ij}^2 - \sum \frac{y_{i.}^2}{n_i}$$

is a value of the random variable F, having an F distribution with $k-2$ and $n-k$ degree of freedom under the null hypothesis that the relationship is linear and therefore may be used to test the hypothesis H_0 for linearity of regression.

When H_0 is true, $\chi_1^2/(k-2)$ and $\chi_2^2/(n-k)$ are independent estimates of σ^2. However, if H_0 is false, $\chi_1^2/(k-2)$ overestimates σ^2. Hence, we reject the hypothesis of linearity of regression at the α level of significance when our f value falls in a critical region of size α located in the upper tail of the F distribution.

Multiple Regression

For the situations with more than one independent variables, $X_1, X_2, \ldots, X_p$, say that are the causes of variation in *Y*, we fit multiple regression of y on $x's$ to account for this variation. Multiple regression of y on $x's$ is denoted as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where β_0 denotes intercept and $\beta_i's$ $(i = 1,2,\ldots,p)$ are called partial regression coefficients. ε is random error. β_i gives average change in y per unit change in x_i keeping other $x's$ constant.

Fitting of Multiple Regression Model

Suppose n observation are made on y and $x's$. Then for each observation we have our unobserved error term ε_i. We make the following assumptions regarding the random variables $\varepsilon_i's$ same as those in simple linear regression case.

In order to estimate the unknown parameters $\beta_0, \beta_1, \beta_{2,}\ldots, \beta_p$, we use the method of least squares which requires minimization of the error sum of squares, given by

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}\right)^2$$

Differentiating S w.r.t. $\beta_0, \beta_1, \beta_{2,}\ldots, \beta_p$ and equating the derivatives to zero, we get a set of $p+1$ equations in $p+1$ unknown as

$$\begin{aligned}
&\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + \ldots + \beta_p \sum_{i=1}^{n} x_{pi} \\
&\sum_{i=1}^{n} x_{1i} y_i = \beta_0 \sum_{i=1}^{n} x_{1i} + \beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \ldots + \beta_p \sum_{i=1}^{n} x_{1i} x_{pi} \\
&\sum_{i=1}^{n} x_{2i} y_i = \beta_0 \sum_{i=1}^{n} x_{2i} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \beta_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + \beta_p \sum_{i=1}^{n} x_{2i} x_{pi} \\
&\vdots \\
&\sum_{i=1}^{n} x_{pi} y_i = \beta_0 \sum_{i=1}^{n} x_{pi} + \beta_1 \sum_{i=1}^{n} x_{1i} x_{pi} + \beta_2 \sum_{i=1}^{n} x_{2i} x_{pi} + \ldots + \beta_p \sum_{i=1}^{n} x_{pi}^2
\end{aligned}$$

These normal equations can be solved simultaneously to get $p+1$ unknowns. However, it is better to solve these equations by inverting the matrix of coefficients of right hand side as this enables us to test significance of $\beta's$ in a straightforward manner. The above equations can be written as

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{pi} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{pi} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{pi} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{pi} & \sum_{i=1}^{n} x_{1i}x_{pi} & \sum_{i=1}^{n} x_{1i}x_{pi} & \cdots & \sum_{i=1}^{n} x_{pi}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{pi}y_i \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{pi} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{pi} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{pi} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{pi} & \sum_{i=1}^{n} x_{1i}x_{pi} & \sum_{i=1}^{n} x_{1i}x_{pi} & \cdots & \sum_{i=1}^{n} x_{pi}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i}y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{pi}y_i \end{bmatrix}$$

Let the inverse of the matrix be denoted by

$$c = \begin{bmatrix} c_{00} & c_{01} & c_{02} & \cdots & c_{0p} \\ & c_{11} & c_{12} & \cdots & c_{1p} \\ & & c_{22} & \cdots & c_{2p} \\ & & & \cdots & \vdots \\ & & & & c_{pp} \end{bmatrix}$$

Then , $Var(b_i) = c_{ii}\sigma^2$, $i = 0,1,2,\cdots p$ and $Cov(b_i, b_j) = c_{ij}\sigma^2$.

Estimation of σ^2

In addition to estimating $\beta's$ an estimate of σ^2 is required to test hypotheses and construct interval estimates pertinent to the regression model. Ideally, we would like this estimate not to depend on the adequacy of the fitted model. This is only possible when there are several observations on y for at least one value of x, or when prior information concerning σ^2 is available. When this approach cannot be used, the estimate of σ^2 is obtained from the residual or error sum of squares.

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The residual sum of square has $(n-p-1)$ degrees of freedom, because $p+1$ degrees are associated with the estimates $\beta's$ involved in obtaining $\hat{y}_i$. Now the expected value of SSE is $E(SSE) = (n-p-1)\sigma^2$ so an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-p-1} = MSE.$$

The quantity MSE is called the error mean square or the residual mean square. The square root of s^2 is sometimes called the standard error of regression, and it has the same units as the response variable y.

The above splitting of the total sum of squares due to $y's$ into two components can be formally put in an Analysis of variance table, as below:

Analysis of Variance for Multiple Linear Regression

Source of Variation	**d.f.**	**S.S.**	**M.S.**
Regression	p	$\sum b_i S_{xiy}$	MSR
Deviation form Regression (Residual)	$n-p-1$	SSE	$s^2 = SSE/(n-p-1) = MSE$
Total (corrected mean)	$n-1$	S_{yy}	

Estimate of σ^2 is this case works out to be

$$s^2 = \frac{S_{yy} - \sum_{i=1}^{n} b_1 Sx_i y}{n-p-1}$$

Test of Significance of β' s

One may be interested in testing the hypotheses $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ for some i. The appropriate test statistic for testing this is

$$t = \frac{|b_i|}{SE(b_1)} = \frac{|b_i|}{\sqrt{c_{ii} s^2}} \text{ as } Var(b_i) = c_{ii} s^2$$

Which follows, under the hypothesis a *t*-distribution with $(n-p-1)$ d.f, if H_0 is true. Reject H_0 if $t > t_{\alpha/2, n-p-1}$.

Multiple correlation coefficient (R)

The correlation coefficient between the observed values y_i and predicted values $\hat{y}_i$ is termed as multiple correlation coefficient (R). Note that $0 \le R \le 1$. R is obtained as

$$R = \sqrt{\frac{\text{Sumof Squaresduetoregression}/\beta_0}{\text{Total corrected sum of squaresof } y}}$$

$$= \sqrt{\frac{b_0 \sum_{i=1}^{n} y_i + b_1 \sum_{i=1}^{n} x_{1i} y_i + b_2 \sum_{i=1}^{n} x_{2i} y_i + b_p \sum_{i=1}^{n} x_{pi} y_i - \left(\sum_{i=1}^{n} y_i \right)^2 \frac{1}{n}}{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y \right)^2 \frac{1}{n}}}$$

$$= \sqrt{\frac{\sum_{i=1}^{p} b_i S_{xiy}}{S_{yy}}}$$

Test of Significance of R

The test of the null hypothesis that multiple correlation coefficient in the population is zero is identical to the F-test of the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. The relation is

$$F = \frac{R^2}{1 - R^2} \quad \frac{n - p - 1}{p}.$$

This F follows F-distribution with p and $(n - p - 1)$ d.f. Reject H_0 if $F > F_{\alpha, p, n-p-1}$.

Coefficient of Determination $\left(R^2\right)$

The sample coefficient of multiple determination, denoted by $R^2_{Y.12\ldots p}$, is given by

$$R^2_{Y.12} \;=\; 1 - \frac{SSE}{(n-1) s_y^{2'}}$$

where $SSE = S_{yy} - b_1 Sx_1 y - b_2 Sx_2 y - \cdots - b_p Sx_p y$

One can easily see that the coefficient of multiple determination is the square of multiple correlation coefficient and is denoted by $\left(R^2\right)$. This concept is very important as $R^2 \times 100$ gives percentage of variation in y explained by regressors. Obviously R^2 must lie between 0 and 1. Thus R^2 is an indicator of fitness of the fitted model. However, a large value of R^2 should not alone be taken as a measure of goodness of fitted regression model.

4. Discussion

In addition to predicting the outcome variable for a new sample of data, regression analysis serves other purposes:

- To assess how well the dependent variable can be explained by knowing the value of the independent variable (or a set of independent variables).
- To identify which subset from many measures is most effective for estimating the dependent variable.

For this, one should first explore the variables graphically in scatterplots to ascertain if a linear model is appropriate for describing the relationship and to identify any possible rogue values (outliers) that might distort results. Ideally, in an observational study, the configuration of plot points should form the shape of an American football, for there are fewer points at the low and high ends of the independent variable than in the middle. In an experimental study, the values of *x* are fixed or set at specified levels, so the configuration may not exhibit such a clear pattern.

In assessing the suitability of the data for a regression, it helps to think of the fixed *x* situation. Visually scan the distribution of *y* values for each *x* (or each small range of *x*'s) – that is, look at vertical strips or bands of points extending up from the x axis. Do the y values within each strip look like a sample from a normal distribution? Is the spread (variance) within each strip roughly the same across the strips? Or is it considerably greater at one side of the plot than at the other? If you guess an average value of y for each strip, do these averages fall along a straight line?

More formally, normality is not required for the estimates of the coefficients. To make tests and estimate confidence intervals, however, these assumptions are required:

- The errors are normally distributed with mean 0.
- The errors have constant variance.
- The errors are independent of each other.

These assumptions are checked by studying the residuals from the model. The **Durbin-Watson** statistic can be used to test for the serial correlation of adjacent error terms.

To identify problems, always look at plots of y versus x before the regression and plots of residuals and diagnostics after the analysis. Non-linearity, Outliers and the presence of sub-populations can distort the results of regression analysis. Relationships among the dependent and independent variables may be *masked or falsely enhanced* if your sample contains subpopulations (that is, the sample is not homogeneous).

In summary, to help identify problems, always look at plots of y versus x before the regression and plots of residuals and diagnostics after the analysis.

Steps for fitting a regression equation using MS-EXCEL: Prepare your data in a Worksheet. Now choose Tools → Data Analysis → Regression. Then give the range for dependent variable, independent variables and output range. If a regression equation without intercept is required then check on Intercept zero.

5. Exercises

Exercise 5.1: The following data are taken from Berenson and Levine (1992). Fifteen similar homes built by one developer in various locations around the United States were evaluated in the study. The builders recorded the amount of oil consumed in January, the average outside temperature (in degree Fahrenheit), and the number of inches of attic insulation in each home.

Case	Avg.Temp.	Insulation(Inches)	Oil Consumed in January
1	40	3	275.3
2	27	3	363.8
3	40	10	164.3
4	73	6	40.8
5	64	6	94.3
6	34	6	230.9
7	9	6	366.7
8	8	10	300.6
9	23	10	237.8
10	63	3	121.4
11	65	10	31.4
12	41	6	203.5
13	21	3	441.1
14	38	3	323.0
15	58	10	52.5

1. Draw all possible scatter plots matrix by taking two variables at a time.
2. Fit a multiple linear regression equation using oil consumed as dependent variable and insulation and average temperature as independent variable

Exercise 5.2: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (Kg./plot).

1. Plot a simple scatter diagram between (i) yield and PP (ii) yield and PH (iii) yield and NGL.
2. Compute bivariate and partial correlations among yield, PP, PH and NGL.
3. Fit a multiple linear regression by taking yield as dependent variable.

No.	PP	PH	NGL	Yield
1	142.00	0.5250	8.20	2.470
2	143.00	0.6400	9.50	4.760
3	107.00	0.6600	9.30	3.310
4	78.00	0.6600	7.50	1.970
5	100.00	0.4600	5.90	1.340
6	86.50	0.3450	6.40	1.140
7	103.50	0.8600	6.40	1.500
8	155.99	0.3300	7.50	2.030
9	80.88	0.2850	8.40	2.540
10	109.77	0.5900	10.60	4.900
11	61.77	0.2650	8.30	2.910
12	79.11	0.6600	11.60	2.760
13	155.99	0.4200	8.10	0.590
14	61.81	0.3400	9.40	0.840
15	74.50	0.6300	8.40	3.870

16	97.00	0.7050	7.20	4.470
17	93.14	0.6800	6.40	3.310
18	37.43	0.6650	8.40	1.570
19	36.44	0.2750	7.40	0.530
20	51.00	0.2800	7.40	1.150
21	104.00	0.2800	9.80	1.080
22	49.00	0.4900	4.80	1.830
23	54.66	0.3850	5.50	0.760
24	55.55	0.2650	5.00	0.430
25	88.44	0.9800	5.00	4.080
26	99.55	0.6450	9.60	2.830
27	63.99	0.6350	5.60	2.570
28	101.77	0.2900	8.20	7.420
29	138.66	0.7200	9.90	2.620
30	90.22	0.6300	8.40	2.000
31	76.92	1.2500	7.30	1.990
32	126.22	0.5800	6.90	1.360
33	80.36	0.6050	6.80	0.680
34	150.23	1.1900	8.80	5.360
35	56.50	0.3550	9.70	2.120
36	136.00	0.5900	10.20	4.160
37	144.50	0.6100	9.80	3.120
38	157.33	0.6050	8.80	2.070
39	91.99	0.3800	7.70	1.170
40	121.50	0.5500	7.70	3.620
41	64.50	0.3200	5.70	0.670
42	116.00	0.4550	6.80	3.050
43	77.50	0.7200	11.80	1.700
44	70.43	0.6250	10.00	1.550
45	133.77	0.5350	9.30	3.280
46	89.99	0.4900	9.80	2.690

SAMPLING DISTRIBUTIONS

1. Introduction

The term **population** is referred to any collection of individuals or of their attributes or of results of operations which can be numerically specified. Thus, there may be population of weights of individuals, heights of trees, prices of wheat, number of plants in a field, number of students in a university etc. A population with finite number of individuals or members is called a finite population. For instance, the population of ages of twenty boys in a class is an example of finite population. A population with infinite number of members is known as infinite population. The population of pressures at various points in the atmosphere is an example of infinite population. For any statistical investigation with large population size, **complete enumeration** (or census) of the population is impracticable, for example, estimation of average monthly income of the individuals in the entire country. Further, in some cases, if the population is infinite, then the complete enumeration is impossible. As an illustration, to know the total amount of timber available in the forest, the entire forest cannot be cut to know how much timber is available there.

To overcome the difficulties of complete enumeration, a part or fraction is selected from the population which is called a **sample** and the process of such selection is called sampling. For example, only 20 students are selected from a university or 10 plants are selected from a field. For determining the population characteristic, instead of enumerating all the units in the population, the units in the sample only are observed and the parameters of the population are estimated accordingly. Sampling is therefore resorted to when either it is impossible to enumerate all the units in the whole population or when it is too costly to enumerate in terms of time and money or when the uncertainty inherent in sampling is more than compensated by the possibilities of errors in complete enumeration. The theory of sampling is based on the logic of particular (i.e. sample) to general (i.e. population) and hence all results will have to be expressed in terms of probability. To serve a useful purpose, sampling should be unbiased and representative.

The aim of the theory of sampling is to get as much information as possible, ideally the whole of the information about the population from which the sample has been drawn. In particular, given the form of the parent population, one would like to estimate the parameters of the population or specify the limits within which the population parameters are expected to lie with a specified degree of confidence.

The fundamental assumption underlying most of the theory of sampling is **random sampling** which consists in selecting the individuals from the population in such a way that each individual of the population has the same chance of being selected. Suppose a sample of size n is taken from a finite population of size *N*. Then there would be $\binom{N}{n} = \dfrac{N!}{n!(N-n)!}$, where *n! = n(n-1)…1*, possible samples. A sampling technique in which each of the $\binom{N}{n}$ samples has an equal chance of being selected is known as **simple random sampling**. Some of the other commonly used sampling procedures are: Purposive sampling, stratified sampling, systematic sampling and cluster sampling. The type of sampling to be adopted depends on the objective of the study and the variability in the population. It may be pointed out that throughout the term random sampling would always refer to simple random sampling.

1.1 Definitions of Some Important Concepts

Parameter: A **parameter** is a function of population values.

Example 1.1: Suppose X_1, X_2,..., X_N are N population values. Then population mean (μ) is a parameter defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \, .$$

Further, population standard deviation (σ) is another parameter defined as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

Statistic: A **statistic** is a function of sample values.

Example 1.2: Two of the most commonly used statistics based on a sample of size n are sample mean ($\bar{x}$) and sample standard deviation (s) defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \, ,$$

$$s = \sqrt{\frac{1}{n\text{-}1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

In practice, parameter values are not known and their estimates based on the sample values are used. A statistic is since based on sample values, there can be many choices of the samples that can be drawn from the population. The distribution of the statistic computed for all possible values of the sample is called **sampling distribution**. From the given set of observations, different statistics are constructed to estimate the parameters. The sampling distributions of these statistics will, in general, depend on the form and the parameters of the parent population. The probability of the observed value of the statistics then allows the making of statements about the parameters and hence conclusions can be drawn about the population. The sampling distributions are thus fundamental to the entire subject of inference and are described below.

2. Sampling Distribution Based on Other Statistics

Distribution based on some statistics when random sample has been drawn from a normal population are now described.

2.1 Sampling Distribution of Sample Mean

If x_1, x_2,..., x_n is a random sample from a normal distribution with mean μ and variance σ^2, then $\bar{x}$ as defined in Example 1.2 follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Further, $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ follows a normal distribution with mean 0 and variance 1, i.e. N(0,1).

Let Z have a standard normal distribution, the probability that Z will exceed a given value z is α, i.e.

$$P[Z > z_\alpha] = \alpha, \quad 0 \leq \alpha \leq 1$$

In other words, as shown by shaded portion in the figure below, α is "Area to the right of the point z_α". The probabilities for negative values of z can be obtained by symmetry

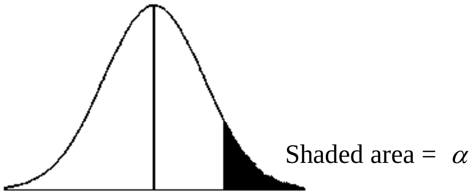


Table 1 gives the normal probability (right tail area) for various values of z.

Example 2.1: Let $\bar{x}$ be the mean of a random sample of size 5 from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 125$, then

$$P\{\bar{x} > 10\} = P\{\frac{\bar{x}-0}{\sqrt{125/5}} > \frac{10-0}{5}\} = P\{Z > 2\} = 0.0228.$$

2.2 Chi-Square Distribution

This distribution was initially proposed by F.R. Helmert but later on also given independently by Karl Pearson. It is defined as the distribution of sum of squares of n independent standard normal variates i.e. if Z_1, Z_2, ...Z_n are n independent standard normal variates, then $\sum_{i=1}^{n} Z_i^2$ follows χ^2 distribution with n degrees of freedom and symbolically it can be written as χ_n^2.

It is to be noted here that, square of a standard normal variate will follow a χ^2 distribution with 1 degrees of freedom.

Thus, If X follows χ^2 distribution with n degrees of freedom, then the probability density function (pdf) is

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{(n/2)-1}, \qquad 0 \leq x < \infty$$

The parameter n is also called **degrees of freedom**, which is a measure of the number of independent variables.

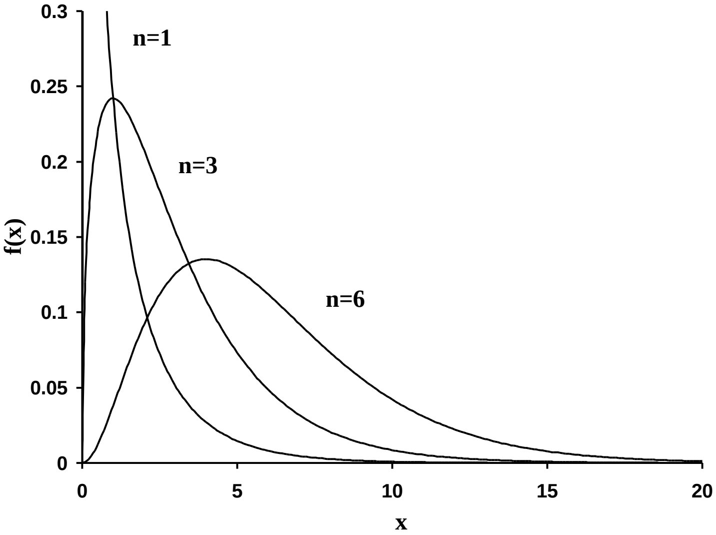
Properties

- Mean = n
- Variance = $2n$
- Mode = n-2, if $n > 2$

- M.G.F.: $(1-2t)^{-\frac{n}{2}}$, $|2t|<1$
- Thus the distribution is positively skewed and leptokurtic
- Sum of independent χ^2-variates is a χ^2-variate (additive property)

Graph

Given below is the graph of chi-square distribution for degrees of freedom n = 1, 3 and 6. In case of n = 1, the mode does not exist. It is seen that the graphs are to the right of 0 and flattens out towards the right (positively skewed). Also there is no symmetry in the graph.



Table

Let X have a χ^2 distribution with n degrees of freedom. Then $\chi^2_n(\alpha)$ is defined as that value which satisfies

$$P[X > \chi^2_n(\alpha)] = \alpha, \quad 0 \le \alpha \le 1$$

In other words, as shown by shaded portion in the figure below, α is "Area to the right of the point $\chi^2_n(\alpha)$".

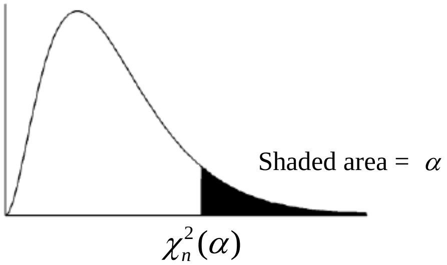


Table 2 gives values of $\chi^2_n(\alpha)$ for various values of α and n.

Example 2.2: Let *X* have a chi-square distribution with 7 degrees of freedom. Then from Table 2,

(i) $\chi^2_7(0.05) = 14.067$

(ii) $\chi^2_7(0.95) = 2.167$.

Application

- To test hypothetical value of population variance
- Goodness-of-fit Test
- Independence of Attributes Test
- To test homogeneity of independent estimates of population variance (Bartlett's Test)
- To test homogeneity of independent estimates of population correlation coefficient

2.3 Sampling Distribution of Sample Variance

Suppose random samples of size n are drawn repeatedly from a normal population with variance σ^2, then

$$X = \frac{(n-1)s^2}{\sigma^2},$$

follows a χ^2_{n-1} distribution. Thus, sample variance s^2 follows $\frac{\sigma^2}{(n-1)}\chi^2_{n-1}$. Mean of s^2 is σ^2 and variance is $\frac{2\sigma^4}{(n-1)}$.

2.4 t - Distribution

This distribution was proposed by Sir R.A. Fisher, who is known as the Father of Statistics. It is defined as the distribution of the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom.

In other words, let Z be a standard normal variate, X a χ^2_n variate and if Z and X are independent, then

$$t = \frac{Z}{\sqrt{X/n}}$$

has a t - distribution with n degrees of freedom and its pdf is

$$f(t) = \frac{1}{\sqrt{n}\,B(\frac{1}{2},\frac{n}{2})}\frac{1}{[1+\frac{t^2}{n}]^{(n+1)/2}}, \qquad -\infty < t < \infty$$

where $B(l, m)$ is beta function defined as $B(l, m) = x^{(l-1)}(1-x)^{m-1}$, $(l, m) > 0$.

If $x_1, x_2, \ldots, x_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then the random variable $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ is distributed as Student's t-distribution with n-1 degrees of freedom

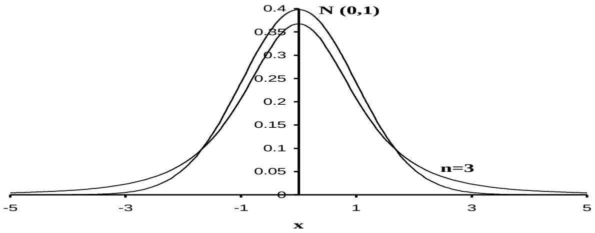
Properties

- Mean = 0, if n > 1
- Variance = $\frac{n}{n-2}$, n > 2

- Mode = 0
- As, t-distribution is symmetric, all odd order moments about zero as well as all odd order central moments are zero.
- t- distribution leptokurtic
- M.G.F. does not exist for t-distribution
- t-distribution tends to normal distribution as n tends to infinity

Graph

The graph below shows the probability density function of t-distribution with 3 degrees of freedom along with that of standard normal distribution N(0,1). The graph of t-distribution is symmetrical with respect to vertical axis x = 0. It is a bell shaped curve and the spread increases as n decreases.



Table

Let $t_n(\alpha)$ denote the point for which

$$P\,[X > t_n(\alpha)] = \alpha,$$

In other words, as shown by shaded portion in the figure below, α is “Area to the right of the point $t_n(\alpha)$”.

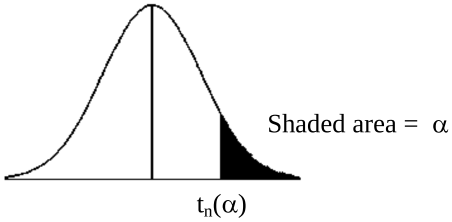
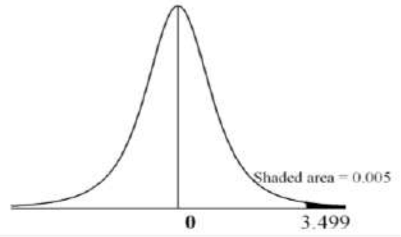


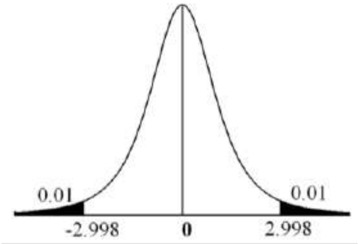
Table 3 gives values of $t_n(\alpha)$ for various values of α and n.

Example 2.3: Let *X* have a t-distribution with 7 degrees of freedom, then from Table 3,

(i) $P(X>3.499) = 0.005$ or $t_7(0.005) = 3.499$



(ii) $P(X \le -2.998) = P(X > 2.998) = 0.01$



Applications

- t-test for single Mean
- t-test for equality of means for two independent sample
- t-test for equality of means for paired observation i.e. paired t-test
- t-test for significance of an observed sample correlation coefficient
- t-test for significance of an observed sample regression coefficient
- t-test for significance of an observed partial regression coefficient

2.5 F-Distribution

The F distribution was discovered by George Snedecor, but in honour of Sir R.A. Fisher, he called it F-distribution. The distribution is defined as the distribution of the ratio of two independent chi-square variates, divided by their respective degrees of freedom, i.e. if X follows $\chi^2_{n_1}$, Y follows $\chi^2_{n_2}$ and X and Y are independent,

$$F = \frac{X/n_1}{Y/n_2},$$

follows F-distribution with (n_1, n_2) degrees of freedom and its pdf is

$$f(x) = \frac{(n_1/n_2)^{n_1/2}}{B(\frac{n_1}{2},\frac{n_2}{2})} \frac{x^{(n_1/2)-1}}{(1+\frac{n_1}{n_2}x)^{(n_1+n_2)/2}}, \qquad 0 \le x < \infty$$

If $x_1, x_2, ..., x_{n_1}$ is a random sample of size n_1 from a normal distribution with mean μ_x and variance σ_x^2 and $y_1, y_2, ..., y_{n_2}$ is a random sample of size n_2 from an independent normal distribution with mean μ_y and σ_y^2, then random variable $F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ has a Snedecor's F-distribution with n_1-1 and n_2-1degrees of freedom.

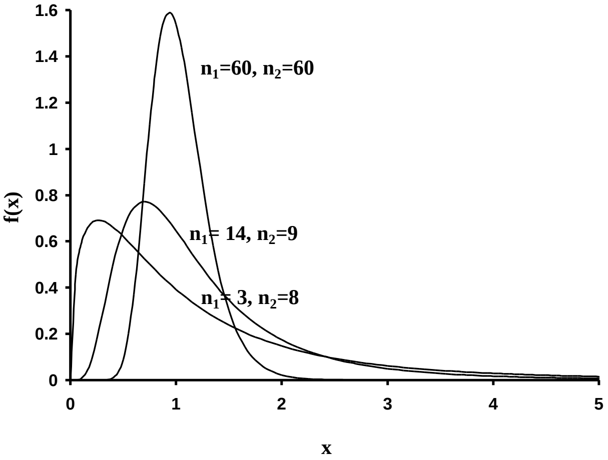
Properties

- Mean = $\frac{n_2}{n_2-2}$, $n_2 > 2$
- Variance = $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$, $n_2 > 4$
- Mode = $\frac{n_2(n_1-2)}{n_1(n_2+2)}$, $n_1 > 2$

  Thus mode of F_{n_1,n_2} distribution, whenever existent, is less than unity.
- There exists a reciprocal relation, i.e. $F_{n_1,n_2}(\alpha) = \frac{1}{F_{n_2,n_1}(1-\alpha)}$
- F_{n_1,n_2} distribution tends to normal distribution when n_1 and n_2 tend to infinity

Graph

The curve of F distribution depends not only on the two parameters n_1 and n_2 but also on the order in which these occur. It is positively skewed.



Table

Let $F_{n_1,n_2}(\alpha)$ denote the point for which

$$P[X > F_{n_1,n_2}(\alpha)] = \alpha,$$

In other words, as shown by shaded portion in the figure below, α is "Area to the right of the point $F_{n_1,n_2}(\alpha)$".

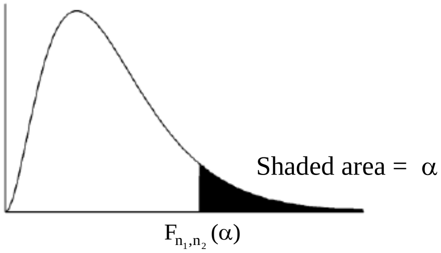
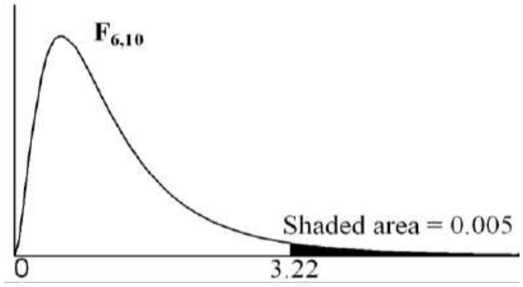


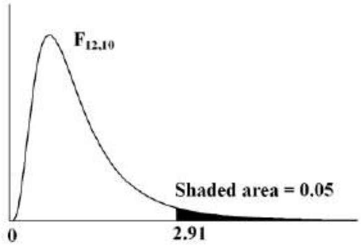
Table 4 and Table 5 give the values of $F_{n_1,n_2}(\alpha)$ for α = 0.01 and 0.05 respectively for various combinations of the degrees of freedom n_1 and n_2.

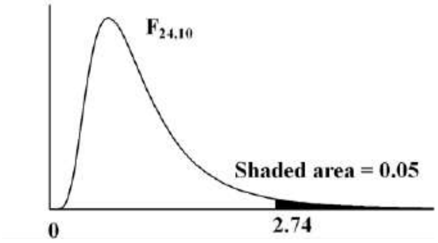
Example 2.4: The value of F with 6 and 10 degrees of freedom, leaving an area of 0.05 to the right, is $F_{6,10}(0.05) = 3.22$.

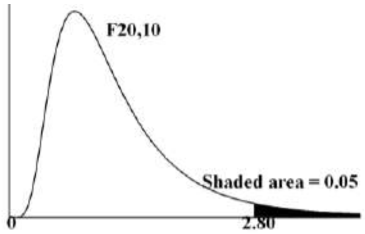


Example 2.5: Given $F_{4,5}(0.05) = 5.19$, then $F_{5,4}(0.95) = \frac{1}{5.19} = 0.19$.

Example 4.5.3: Given $F_{12,10}(0.05) = 2.91$and $F_{24,10}(0.05) = 2.74$, then $F_{20,10}(0.05) = 2.80$.







Example 4.5.4: Given $F_{6,30}(0.05) = 2.42$ and $F_{6,40}(0.05) = 2.34$, then $F_{6,35}(0.05) = 2.38$.

2.6 Relation Between t, χ^2 and F distributions

(i) If a random variable X follows t-distribution with n degrees of freedom, then X^2 follows F-distribution with n_1 =1 and n_2 = n degrees of freedom. In other words, $t_n^2(\frac{\alpha}{2}) = F_{1,n}(\alpha)$.

(ii) If a random variable X follows F-distribution with n_1 and n_2 degrees of freedom, then as n_2 tends to infinity, n_1X follows Chi-square with n_1 degrees of freedom.

Table 1: Percentage points (right tail area) of normal distribution for various values of z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.2272	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2231	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1984	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0017	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Table 2: Percentage points (right tail area) of chi-square distribution for various values of α and n

α n	0.99	0.975	0.95	0.50	0.10	0.05	0.025	0.01
1	0.000	0.001	0.004	0.455	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	1.386	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	2.366	6.251	7.815	9.348	11.341
4	0.297	0.484	0.711	3.357	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	4.351	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	5.348	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	6.346	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	7.344	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	8.343	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	9.342	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	10.341	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	11.340	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	12.340	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	13.339	21.064	23.635	26.119	29.141
15	5.229	6.262	7.261	14.339	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	15.338	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	16.338	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	17.338	25.989	28.869	31.526	34.802
19	7.633	8.907	10.117	18.338	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	19.337	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	20.337	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	21.337	30.813	33.924	36.781	40.289
23	10.196	11.688	13.091	22.337	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	23.337	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	24.337	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	25.336	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	26.336	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	27.336	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	28.336	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	29.336	40.256	43.773	46.979	50.892

Table 3: Percentage points (Two tail areas) of t-distribution for various values of α and n

α n	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
35	0.682	1.307	1.690	2.029	2.440	2.720
40	0.681	1.303	1.684	2.021	2.423	2.704
45	0.680	1.302	1.683	2.020	2.410	2.690
50	0.679	1.298	1.674	2.010	2.400	2.680
60	0.679	1.296	1.671	2.000	2.390	2.660
∞	0.674	1.282	1.645	1.960	2.326	2.576

Table 4: Percentage points (right tail area) of F-distribution for various values of n_1 and n_2 for $\alpha = 0.01$

n_1 n_2	1	2	3	4	5	6	8	12	24	∞
1	4052.00	4999.50	5403.00	5625.00	5764.00	5859.00	5982.00	6106.00	6234.00	6366.00
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.11
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.01
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.95	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.24	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.10	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.91	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.11	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.00	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.30	2.12	1.60
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Table 5: Percentage points (right tail area) of F-distribution for various values of n_1 and n_2 for $\alpha = 0.05$

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	238.90	243.90	249.00	254.30
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.96	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28.	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.31	2.13	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4..08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4..00	3.15	2.76	2.52	2.37	2.25	2.10	1.92.	1.70	1.39
∞	3.84	2.99	2.60	2.37	2.26	2.10	1.94	1.75	1.52	1.00

TESTING OF HYPOTHESIS

1. Introduction

In applied investigations, one is often interested in comparing some characteristic (such as mean or variance) of a group with a specified value, or in comparing two or more groups with regard to the characteristic. For instance, one may want to know whether mean timber yield obtained from recently felled plantations of a particular age in a particular management unit is some specifid value, one may wish to know whether average yield of a crop in a certain district is equal to a specified value, one may wish to compare two species of trees with regard to mean height, to know if genetic fraction of total variation in a strain is more than a given value. In making such comparisons, one can not rely on mere numerical magnitudes of index of comparison such as mean and variance. This is because each group is represented only by a sample of observations and if another sample were drawn, the numerical value would change. This variation between samples from the same population can at best be reduced in a well-designed controlled experiment but can never be eliminated. One is forced to draw inferences in presence of sampling fluctuations which affect observed differences between groups, clouding real differences. Statistical science provides an objective procedure for distinguishing whether observed difference connotes any real difference among groups. Such a procedure is called **testing of hypothesis**. Thus, in short, testing of hypothesis is a method of making due allowance for sampling fluctuation affecting results of experiments or observations. These tests have wide applications in agriculture, forestry, medicine, industry, social sciences, etc.

1.1 Definitions

Statistical Hypothesis: It is an assumption either about the form or about the parameters of a distribution. For example, average height of a particular species of tree is 50 feet, normal distribution has mean 20.

If all the parameters are completely specified, hypothesis is called a **simple hypothesis**, otherwise it is a **composite hypothesis**. For example, average height of tree is 50 feet is a simple hypothesis and average height of tree is greater than 50 feet is a composite hypothesis.

Null Hypothesis (H_0): The hypothesis under test for a sample study is called Null hypothesis (H_0). It represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, null hypothesis might be that the new drug is, on average, as effective as the current drug i.e. H_0: Effect of the two drugs, on the average, is same.

Alternative Hypothesis (H_1): Any null hypothesis is tested against a rival, which is called Alternative hypothesis (H_1). For example, mean height (μ) of trees of a particular species in a region is some specified value μ_0, i.e.

H_0: $\mu = \mu_0$.

Alternative hypothesis could be any of the following:

H_1: $\mu \neq \mu_0$ (Two-tailed)
$\mu < \mu_0$ (Left-tailed)
$\mu > \mu_0$ (Right-tailed)

For framing a suitable H_0 and H_1, four possibilities in order of preference are the following:

Possibilities	**H_0**	**H_1**
(i)	Simple	Simple
(ii)	Simple	Composite
(iii)	Composite	Simple
(iv)	Composite	Composite

The first one when both are simple is of little practical importance. As Possibility (ii) is preferred over Possibility (iii), therefore hypotheses should always be structured in such a way that H_0 is simple and H_1 is composite.

Two Types of Errors

True Situation → **Decision Made ↓**	**H_0 is True**	**H_0 is False**
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

Probabilities of these types of error are respectively denoted by α and β, i.e.

Probability of Type I error = α

and Probability of Type II error = β.

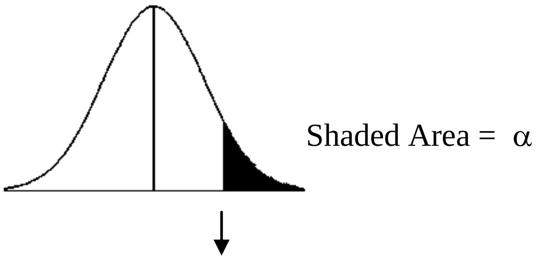
The ideal procedure of hypothesis testing is to minimize both α and β. However, this is not possible in practice because a test which minimizes one type of error, maximizes the other type of error. As Type I error is considered to be more serious than Type II error, therefore probability of Type I error is fixed and probability of Type II error is minimized. Generally, α is taken to be 5% or 1%.

Level of Significance (α): It is the size of Type I error. The higher the value of α, less precise is the result.

Confidence Interval: The confidence interval of a parameter with confidence coefficient 100(1-α)% is the interval (a, b) such that it is expected to lie in this interval in 100(1-α)% cases.

Test Statistic: A test statistic is a quantity calculated from data. Its value is used to decide whether or not the null hypothesis should be rejected.

Critical Value(s): The critical value(s) is that value with which value of test statistic in a sample is compared to determine whether or not the null hypothesis is rejected. The critical value for any hypothesis test depends on significance level α at which the test is carried out, and whether the test is one-sided or two-sided.



Critical Value (one-sided)

Power of a Test: It is defined as the probability of rejecting H_0 when it is false. Thus,

$$\text{Power} = 1 - \beta$$

Among a given set of tests, best test is one having maximum power.

Steps in Hypothesis Testing

- State statistical hypotheses
- Check assumptions
- Calculate test statistic
- Set the test criteria
- Interpret the results

We now discuss some tests of hypothesis that are based on normal, t, F and chi-square distributions.

2. Test of Significance for Large Samples

For large n (sample size), almost all the distributions can be approximated very closely by a normal probability curve, we therefore use the **normal test** of significance for large samples. If t is any statistic (function of sample values), then for large sample

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} = N(0.1)$$

Thus if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than Z_α times the standard error (S.E), hypothesis is rejected at α level of significance. Similarly if

$$\left| t - E(t) \right| \leq Z_\alpha \times S.E(t),$$

the deviation is not regarded significant at 5% level of significance. In other words the deviation t - E(t), could have arisen due to fluctuations of sampling and the data do not provide any evidence against the null hypothesis which may, therefore be accepted at α level of significance.

If $|Z| \leq 1.96$, then the hypothesis H_0 is accepted at 5% level of significance. Thus the steps to be used in the normal test are as follows:

i) Compute the test statistic Z under H_0.

ii) If $|Z| > 3$, H_0 is always rejected

iii) If $|Z| < 3$, we test its significance at certain level of significance

The table below gives some critical values of Z:

Level of Significance	**Critical Value (Z_α) of Z**	
	Two-tailed test	**Single tailed test**
10%	1.645	1.280
5%	1.960	1.645
1%	2.580	2.330

2.1 Test for Single Mean

A very important assumption underlying the tests of significance for variables is that the sample mean is asymptotically normally distributed even if the parent population from which the sample is drawn is not normal.

If x_i (i =1,…,n) is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$. Based on this random sample, our aim is to test that mean of the population has a specified value μ_0, i.e.

H_0: $\mu = \mu_0$

The alternative hypothesis could be any of the following:

H_1: $\mu \neq \mu_0$ (two tailed)
$\mu < \mu_0$ (left tailed)
$\mu > \mu_0$ (right tailed)

Test Statistic:

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

follows a standard normal distribution.

Test Criteria: Depending on the alternative hypothesis selected, the test criteria are as follows:

H_1	**Test**	**Reject H_0 at level of significance α if**
$\mu \neq \mu_0$	Two-tailed	$\lvert Z \rvert > Z_{\alpha/2}$
$\mu < \mu_0$	Left-tailed	$Z < -Z_\alpha$
$\mu > \mu_0$	Right-tailed	$Z > Z_\alpha$

Z_α is the table value of Z at level of significance α. If σ^2 is unknown, then it is estimated by sample variance s^2 (for large n), where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Example 2.1: The mean timber yield obtained from 30 recently felled plantations at the age of 50 years in a particular management unit is 93 m^3/ha with a standard deviation of 10 m^3/ha. Test whether the mean timber yield is 100 m^3/ha based on past records.

Solution: H_0 : μ = 100 m^3/ha, H_1: $\mu \neq$ 100 m^3/ha (two tailed test).

Here, $\overline{x} = 93$ m^3/ha., n = 30, μ = 100 m^3/ha and σ = 10 m^3/ha.

Thus,

$$Z = \frac{93-100}{10/\sqrt{30}} = -3.834$$

Since $|Z| > 1.96$, we conclude that the data does not provide any evidence in favour of the null hypothesis H_0 may therefore be rejected at 5% level of significance. Hence the decision would be to accept the alternative hypothesis that there has been significant decline in the productivity of the management unit with respect to the plantations of the species considered.

Note: The value of sample mean is an acceptable value of population mean if the statistic Z lies between $-Z_{\alpha/2}$ to $Z_{\alpha/2}$, i.e.

$$-Z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}.$$

Thus, 100(1-α)% confidence-interval for μ is

$$(\bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{x} + Z_{\alpha/2}\,\sigma/\sqrt{n}).$$

2.2 Test for Difference of Means

Let $\bar{x}_1\,(\bar{x}_2)$ be the mean of a sample of size n_1 (n_2) from a population with mean μ_1 (μ_2) and variance $\sigma_1^2\,(\sigma_2^2)$. Our aim is to test

H_0 : $\mu_1 = \mu_2$

against H_1 : $\mu_1 \neq \mu_2$
$\mu_1 > \mu_2$
$\mu_1 < \mu_2$

Test Statistic:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

follows a standard normal distribution

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$|Z| > Z_{\alpha/2}$
$\mu_1 < \mu_2$	Left-tailed	$Z < -Z_\alpha$
$\mu_1 > \mu_2$	Right-tailed	$Z > Z_\alpha$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{, If } \sigma_1^2 = \sigma_2^2 = \sigma^2$$

If σ is not known, then its estimate is used

$$\hat{\sigma}^2 = s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

2.3 Test for Single Proportion

Suppose in a sample of size n (>30), x be the number of successes. Then observed proportion of successes = $x/n = p$. Let P be the population proportion. The hypothesis to be tested is that population proportion is some specified value P_0, i.e.

H_0: P = P_0
H_1: P ≠ P_0
P > P_0
P < P_0

Test Statistic:

$$Z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}}$$

follows approximately a standard normal distribution.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$P \neq P_0$	Two-tailed	$\lvert Z \rvert > Z_{\alpha/2}$
$P < P_0$	Left-tailed	$Z < -Z_\alpha$
$P > P_0$	Right-tailed	$Z > Z_\alpha$

Example 2.2: In a sample of 1000 people, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular at 1% level of significance?

Solution: It is given that n = 1000, x = Number of rice eaters = 540, p = sample proportion of rice eaters $= 540/1000 = 0.54$.

H_0 : Both rice and wheat are equally popular, i.e. P = 0.5
H_1 : P ≠ 0.5

$$Z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}} = \frac{0.54 - 0.5}{\sqrt{0.5 \text{ x } 0.5/1000}} = 2.532$$

Tabulated value of Z at 1% level of significance is 2.575. Since $\lvert Z \rvert < 2.575$, therefore H_0 is not rejected and we conclude that rice and wheat are equally popular.

2.4 Test for Difference of Proportions

Suppose we want to compare two populations with respect to the prevalence of a certain attribute A. Let x_1 (x_2) be the number of persons possessing the given attribute A in random sample of size n_1 (n_2) from 1st (2nd) population. Then sample proportions will be

$$p_1 = \frac{x_1}{n_1}, \; p_2 = \frac{x_2}{n_2}$$

Let P_1 and P_2 be the population proportions. Our aim here is to test that there is no significant difference between population proportions, i.e.

H_0: $P_1 = P_2$

H_1: $P_1 \neq P_2$

$P_1 > P_2$

$P_1 < P_2$

Test Statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2})}}$$

follows approximately a standard normal distribution. In case $P_1 = P_2 = P$ (say) and P is not known, it is estimated as follows:

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$P_1 \neq P_2$	Two-tailed	$\lvert Z \rvert > Z_{\alpha/2}$
$P_1 < P_2$	Left-tailed	$Z < -Z_\alpha$
$P_1 > P_2$	Right-tailed	$Z > Z_\alpha$

Consider an experiment on rooting of stem cuttings of *Casuarina equisetifolia* wherein the effect of dipping the cuttings in solutions of IBA at two different concentrations was observed. Two batches of 30 cuttings each, were subjected dipping treatment at concentrations of 50 and 100 ppm of IBA solutions respectively. Based on the observations on number of cuttings rooted in each batch of 30 cuttings, the following proportions of rooted cuttings under each concentration were obtained. At 50 ppm, the proportion of rooted cuttings was 0.5 and at 100 ppm, the proportion was 0.37. Test whether the observed proportions are indicative of significant differences in the effect of IBA at the two concentrations.

Here, $p_1 = 0.5$ and $p_2 = 0.37$. Then $q_1 = 0.5$, $q_2 = 0.63$. The value of $n_1 = n_2 = 30$. Thus,

$$Z = \frac{0.5 - 0.37}{\sqrt{\frac{(0.5)(0.5)}{30} + \frac{(0.37)(0.63)}{30}}} = 1.024$$

Since the calculated value of Z (1.024) is less than the table value (1.96) at 5% level of significance, we can conclude that there is no significant difference between proportion rooted cuttings under the two concentration levels.

3. Test of Significance for Small Samples

In this section, the statistical tests based on t, χ^2 and F are given.

3.1 Tests Based on t-Distribution

3.1.1 Test for an Assumed Population Mean

Suppose a random sample $x_1,..,x_n$ of size n (n≥2) has been drawn from a normal population whose variance σ^2 is unknown. On the basis of this random sample the aim is to test

$H_0 : \quad \mu = \mu_0$

$H_0 : \quad \mu \neq \mu_0$

$\mu > \mu_0$

$\mu < \mu_0$

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

The table giving the value of t required for significance at various levels of probability and for different degrees of freedom are called the t – tables which are given in Statistical Tables by Fisher and Yates. The computed value is compared with the tabulated value at α percent level of significance and at (n-1) degrees of freedom and accordingly the null hypothesis is accepted or rejected.

3.1.2 Test for the Difference of Two Population Means

Let $\bar{x}_1(\bar{x}_2)$ be the sample mean of a sample of size n_1 (n_2) from a population with mean μ_1 (μ_2) and variance of the two population be same σ^2, which is unknown. Our aim is to test

$H_0 : \ \mu_1 = \mu_2$

$H_1 : \ \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Let s_i^2, i =1, 2 be sample variances of the two samples. Then common unknown population variance σ^2 is estimated as

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which follows a t-distribution with $n_1 + n_2$ -2 d.f.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$\lvert t \rvert > t_{n_1+n_2-2}(\alpha/2)$
$\mu_1 < \mu_2$	Left-tailed	$t < -t_{n_1+n_2-2}(\alpha)$
$\mu_1 > \mu_2$	Right-tailed	$t > t_{n_1+n_2-2}(\alpha)$

This test statistic is used under certain assumptions *viz*., (i) The variables involved are continuous (ii) The population from which the samples are drawn follow normal distribution (iii) The samples are drawn independently (iv) The variances of the two populations from which the samples are drawn are homogeneous (equal). The homogeneity of two variances can be tested by using F-test.

Example 3.1: A group of 5 plots treated with nitrogen at 20 kg/ha. yielded 42, 39, 48, 60 and 41 kg whereas second group of 7 plots treated with nitrogen at 40 kg/ha. yielded 38, 42, 56, 64, 68, 69 and 62 kg. Can it be concluded that nitrogen at level 40 kg/ha. increases the yield significantly?

Solution: H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 < \mu_2$

Here, $\bar{x}_1 = 46, \bar{x}_2 = 57,\ s^2 = 121.6$

$$t = \frac{46-57}{\sqrt{121.6(\frac{1}{5}+\frac{1}{7})}} = \text{-}1.7 \sim t_{10}$$

Since |t| < 1.81 (value of t at 5% and 10 d.f), the yield from two doses of nitrogen do not differ significantly.

3.1.3 Paired t-test for Difference of Means

When the two samples are not independent but the sample observations are paired together, then this test is applied. The paired observations are on the same unit or matching units. For example, to know the impact of a new teaching method on the performance of students, the observations, in terms of marks, are collected before and after the new teaching method is implemented. Let (x_i, y_i), i = 1,…,n be the pairs of observations and let $d_i = x_i - y_i$. Our aim is to test

H_0 : $\mu_1 = \mu_2$
H_1 : $\mu_1 \neq \mu_2$
$\mu_1 > \mu_2$
$\mu_1 < \mu_2$

Test Statistic:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

follows t distribution with n-1 d.f., where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$\lvert t \rvert > t_{n-1}(\alpha/2)$
$\mu_1 < \mu_2$	Left-tailed	$t < \text{-}t_{n-1}(\alpha)$
$\mu_1 > \mu_2$	Right-tailed	$t > t_{n-1}(\alpha)$

3.1.4 Test for Significance of Observed Correlation Coefficient

Given a random sample (x_i, y_i) , i = 1,…, n from a bivariate normal population. We want to test the null hypothesis that the population correlation coefficient is zero i.e.

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2},$$

where r is the sample correlation coefficient. H_0 is rejected at level α if

$|t| > t_{n-2}\,(\alpha/2)$

This test can also be used for testing the significance of rank correlation coefficient.

3.2 Test of Significance Based on Chi-Square Distribution

3.2.1 Test for the Variance of a Normal Population

Let $x_1, x_2,\ldots,x_n$ $(n\geq 2)$ be a random sample from a normal population with mean μ and variance σ^2. On the basis of this sample our aim is to test

$H_0 : \sigma^2 = \sigma_0^2$

against $H_1 : \sigma^2 \neq \sigma_0^2$

$\sigma^2 < \sigma_0^2$

$\sigma^2 > \sigma_0^2$

Test Statistic:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma_0}\right)^2$$

follows a chi-square distribution with n d.f. when μ is known, and

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

follows a chi-square distribution with n-1 d.f. when μ is not known.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if	
		μ is known	μ is not known
$\sigma^2 \neq \sigma_0^2$	Two-tailed	$\chi^2 < \chi_n^2(1-\alpha/2)$ or $\chi^2 > \chi_n^2(\alpha/2)$	$\chi^2 < \chi_{n-1}^2(1-\alpha/2)$ or $\chi^2 > \chi_{n-1}^2(\alpha/2)$
$\sigma^2 < \sigma_0^2$	Left-tailed	$\chi^2 < \chi_n^2(1-\alpha)$	$\chi^2 < \chi_{n-1}^2(1-\alpha)$
$\sigma^2 > \sigma_0^2$	Right-tailed	$\chi^2 > \chi_n^2(\alpha)$	$\chi^2 > \chi_{n-1}^2(\alpha)$

Tables are available for χ^2 at different levels of significance and with different degrees of freedom.

3.2.2 Test for Goodness of Fit

A test of wide applicability to numerous problems of significance in frequency data is the χ^2 test of goodness of fit. It is primarily used for testing the discrepancy between the expected and the observed frequency, For instance, one may be interested in testing whether a variable like the height of trees follows normal distribution. A tree breeder may be interested to know whether the observed segregation ratios for a character deviate significantly from the Mendelian ratios. In such situations, we want to test the agreement between the observed and theoretical frequencies. Such a test is called a test of goodness of fit.

H_0 : the fitted distribution is a good fit to the given data
H_1 : not a good fit.

Test statistic: If O_i and E_i, i =1,…,n are respectively the observed and expected frequency of ith class, then the statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n\text{-}r\text{-}1}$$

where r is the number of parameters estimated from the sample, n is the number of classes after pooling. H_0 is rejected at level α if calculated χ^2 > tabulated $\chi^2_{n\text{-}r\text{-}1}$ (α).

Example 3.2: In an F_2 population of chillies, 831 plants with purple and 269 with non-purple chillies were observed. Is this ratio consistent with a single factor ratio of 3:1?

Solution: On the hypothesis of a ratio of 3:1, the frequencies expected in the purple and non-purple classes are 825 and 275 respectively.

	Frequency		
	Observed (O_i)	Expected (E_i)	O_i - E_i
Purpose	831	825	6
Non-purple	269	275	-6

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = 0.17.$$

Here χ^2 is based on one degree of freedom. It is seen from the table that the value of 0.17 for χ^2 with 1 d.f corresponds to a level of probability which lies between 0.5 and 0.7. It is concluded that the result is non-significant.

3.2.3 Test of Independence

Another common use of the χ^2 test is in testing independence of classifications in what are known as contingency tables. When a group of individuals can be classified in two ways, the result of the classification in two ways the results of the classification can be set out as follows:

Contingency table

Class	**A_1**	**A_2**	**A_3**
B_1	n_{11}	n_{21}	n_{31}
B_2	n_{12}	n_{22}	n_{32}
B_3	n_{13}	n_{23}	n_{33}

Such a table giving the simultaneous classification of a body of data in two different ways is called contingency table. If there are r rows and c columns the table is said to be an r x c table.

H_0: the attributes are independent
H_1: they are not independent

Test statistic:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

H_0 is rejected at level α if $\chi^2 > \chi^2_{(r-1)(c-1)}$

3.3 Test of Significance Based on F-Distribution

3.3.1 Test for the Comparison of Two Population Variances

Let x_i , i = 1,…,n_1 and x_j, j=1,…,n_2 be the two random samples of sizes n_1 and n_2 drawn from two independent normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_1, \sigma_2^2)$ respectively. s_1^2 and s_2^2 are the sample variances of the two samples.

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n1} (x_i - \bar{x}_1)^2 \text{and } s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n2} (x_j - \bar{x}_2)^2$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n1} x_i, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n2} x_j$$

$H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic: Assuming $s_1^2 > s_2^2$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1 - 1, n_2 - 1}$$

Tables are available giving the values of F required for significance at different levels of probability and for different degrees of freedom. The computed value of F is compared with the tabulated value and the inference is drawn accordingly.

3.3.2 Test for Homogeneity of Several Population Means

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population i.e. they have the same mean. For Example, 5 fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is obtained. The interest is to find whether effects of these fertilizers on the yields is significantly different or in other words, whether the samples have come from the same normal population. This is done through F-test that uses the technique of Analysis of Variance (ANOVA).

ANOVA is the technique of partitioning the total variability into different known components. It consist in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to assignable factors with the estimate due to chance factor or experimental error. The F statistic used for testing the hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (k>2) is

$$F = \frac{\text{Variation among the sample means}}{\text{Variation within the samples}}$$

Practical on Testing of Hypothesis

1. Independent Samples t-Test

An experiment was conducted to evaluate the effect of inoculation with mycorrhiza on the height growth of seedlings of *Pinus kesiya*. In the experiment, 10 seedlings (Group I) were inoculated with mycorrhiza while another 10 seedlings (Group II) were left without inoculation with the microorganism. Following table gives the height of seedlings obtained under the two groups of seedlings:

Plot	Group I	Group II
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2
7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

Test whether inoculated and uninoculated seedlings are significantly different.

Solution: H_0: Mean of Group I (μ_1) = Mean of Group II (μ_2) and H_1: $\mu_1 \neq \mu_2$

From the given data $\bar{x}_1 = 20.8$, $\bar{x}_2 = 9.61$,

$$s_1^2 = \frac{(23.0)^2 + (17.4)^2 + ... + (18.8)^2 - \frac{(208)^2}{10}}{10-1} = \frac{57.24}{9} = 6.36$$

$$s_2^2 = \frac{(8.5)^2 + (9.6)^2 + ... + (7.4)^2 - \frac{(96.1)^2}{10}}{10-1} = \frac{24.3}{9} = 2.7$$

$$s^2 = \frac{(10-1)(6.36) + (10-1)(2.7)^2}{10+10-2} = \frac{57.24 + 24.43}{18} = 4.537$$

$$t = \frac{20.8 - 9.61}{\sqrt{4.737(\frac{1}{10} + \frac{1}{10})}} = 11.75$$

The computed value of t is compared with the tabular value of t (2.10) at $n_1 + n_2 - 2 = 18$ degrees of freedom. Since the computed value is greater than 2.10 and it is concluded that the populations of inoculated and uninoculated seedlings are significantly different with respect to their mean height.

2. Paired t-Test

The following data pertain to organic carbon content measured at two different layers of a number of soil pits. Test whether the mean carbon content from two layers of soil pit differ or not.

Soil pit	Organic Carbon (%)		
	Layer 1 (x)	Layer 2 (y)	Difference (d)
1	1.59	1.21	0.38
2	1.39	0.92	0.47
3	1.64	1.31	0.33
4	1.17	1.52	-0.35
5	1.27	1.62	-0.35
6	1.58	0.91	0.67
7	1.64	1.23	0.41
8	1.53	1.21	0.32
9	1.21	1.58	-0.37
10	1.48	1.18	0.30

The observations are paired by soil pits. The paired t-test can be used in this case to compare the organic carbon status of soil at the two depth levels.

Solution: Mean of Layer 1 (μ_1) = Mean of Layer 2 (μ_2) and H_1: $\mu_1 \neq \mu_2$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{1.81}{10} = 0.181$$

$$s_d^2 = \frac{1}{10-1}\left([(0.38)^2 + (0.47)^2 + ... + (0.30)^2] - \frac{(1.81)^2}{10} \right) = \frac{1.3379}{9} = 0.1486$$

Thus,

$$t = \frac{0.181}{\sqrt{\frac{0.1486}{10}}} = 1.485$$

The value of t (1.485) is less than the tabular value, 2.262, for 9 degrees of freedom at the 5% level of significance. It may therefore be concluded that there is no significant difference between the mean organic carbon content of the two layers of soil.

3. Goodness of Fit

In an F_2 population of chillies, 831 plants with purple and 269 with non-purple chillies were observed. Is this ratio consistent with a single factor ratio of 3:1?

Solution: On the hypothesis of a ratio of 3:1, the frequencies expected in the purple and non-purple classes are 825 and 275 respectively.

	Frequency		
	Observed (O_i)	Expected (E_i)	O_i - E_i
Purpose	831	825	6
Non-purple	269	275	-6

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = 0.17.$$

Here χ^2 is based on one degree of freedom. It is seen from the table that the value of 0.17 for χ^2 with 1 d.f corresponds to a level of probability which lies between 0.5 and 0.7. It is concluded that the result is non-significant.

4. Equality of Several Means (Analysis of Variance)

Ten varieties of wheat are grown in 3 plots each and the following yields in kg per hectare are obtained:

Variety → Plots ↓	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7	7	14	11	9	6	9	8	12	9
2	8	9	13	10	9	7	13	13	11	11
3	7	6	16	11	12	5	11	11	11	11

Test the significance between mean variety yields.

PLANNING AND DESIGNING OF AGRICULTURAL EXPERIMENTS

An experiment is usually associated with a scientific method for testing certain phenomena. An experiment facilitates the study of such phenomena under controlled conditions and thus creating controlled condition is an essential component. Scientists in the biological fields who are involved in research constantly face problems associated with planning, designing and conducting experiments. Basic familiarity and understanding of statistical methods that deal with issues of concern would be helpful in many ways. Researchers who collect data and then look for a statistical technique that would provide valid results will find that there may not be solutions to the problem and that the problem could have been avoided first by a properly designed experiment. Obviously it is important to keep in mind that we cannot draw valid conclusions from poorly planned experiments. Second, the time and cost involved in many experiments are enormous and a poorly designed experiment increases such costs in time and resources. For example, an agronomist who carries out fertilizer experiment knows the time limitation of the experiment. He knows that when seeds are to be planted and harvested. The experimenter plot must include all components of a complete design. Otherwise what is omitted from the experiment will have to be carried out in subsequent trials in the next cropping season or next year. The additional time and expenditure could be minimized by a properly planned experiment that will produce valid results as efficiently as possible. Good experimental designs are products of the technical knowledge of one's field, an understanding of statistical techniques and skill in designing experiments.

Any research endeavor may entail the phases of Conception, Design, Data collection, Analysis and Dissemination. Statistical methodologies can be used to conduct better scientific experiments if they are incorporated into entire scientific process, i.e., from inception of the problem to experimental design, data analysis and interpretation. When planning experiments we must keep in mind that large uncontrolled variations are common occurrences. Experiments are generally undertaken by researchers to compare effects of several conditions on some phenomena or in discovering an unknown effect of particular process. An experiment facilitates the study of such phenomena under controlled conditions. Therefore the creation of controlled condition is the most essential characteristic of experimentation. How we formulate our questions and hypotheses are critical to the experimental procedure that will follow. For example, a crop scientist who plants the same variety of a crop in a field may find variations in yield that are due to periodic variations across a field or to some other factors that the experimenter has no control over. The methodologies used in designing experiments will separate with confidence and accuracy a varietal difference of crops from the uncontrolled variations.

The different concepts in planning of experiment can be well explained through chapati tasting experiment.

Consider an experiment to detect the taste difference in chapati made of wheat flour of c306 and pv 18 varieties. The null hypothesis we can assume here is that there is no taste difference in chapatis made of c306 or pv18 wheat flours. After the null hypothesis is set, we have to fix the level of significance at which we can operate. The pv18 is a much higher yielding variety than c306. Hence a false rejection may not help the country to grow more pv18 and the wheat production may decrease while a false acceptance may give more production of pv18 wheat and the consumption may be less or practically nil. Thus the false acceptance or false rejection are of practically equal consequence and we agree to choose the level of significance at $\alpha = 0.05$. Now to execute the experiment, a subject is to be found with extrasensory powers who can detect the taste differences. The colours of c306 and pv18 are different and anyone, even without tasting the chapatis, can distinguish the chapatis of either kind by a mere glance. Thus the taster of the chapatis has to be blindfolded before the chapatis are given for tasting. Afterwards, the method is to be decided in which the experiment will be conducted. The experiment can be conducted in many ways and of them three methods are discussed here:

- Give the taster equal number of chapatis of either kind informing the taster about it.
- Give the taster pairs of chapatis of each kind informing the taster about it.
- Give the taster chapatis of either kind without providing him with any information. Let us use 6 chapatis in each of these methods.

Under first method of experimentation, if the null hypothesis is true, then the experimenter cannot distinguish the two kinds of chapaties and he will randomly select 3 chapatiS out of 6 chapaties given to him, as made of pvl8 wheat. In that case, all correct guesses are made if selection exactly coincides with the exactly used wheat variety and the probability for such an occurrence is:

$$1\Big/\binom{6}{3} = 1\Big/20 = 0.05$$

Under second method,the pv18 wheat variety chapaties are selected from each pair given if the null hypothesis is true. Furthermore, independent choices are made of pv18 variety chapaties from each pair. Thus the probability of making all correct guesses is

$$1/(2)^3 = 1/8 = 0.125.$$

In third method the experimenter has to make the choice for each chapati and the situation is analogous at calling heads or tails in a coin tossing experiment. The probability of making all correct guesses would then be:

$$1/2^6 = 1/64 = .016.$$

If the experimenter makes all correct guesses in third method as its probability is smaller than the selected $\alpha = 0.05$, we can reject the null hypothesis and conclude that the two wheat varieties give different tastes at chapaties. In other methods the probability of making all correct guesses does not exceed $\alpha = 0.05$ and hence with either method, we cannot reject the null hypothesis even if all correct guesses are made.

However, if 8 chapaties are used by first method and if the taster guesses all of them, we can reject the null hypothesis, at 0.05 level of significance, as the probability of making all correct guesses would then be $1\Big/\binom{8}{3} = 1\Big/56$ which is smaller than 0.05. 8 chapaties will not enable us to reject the null hypothesis even if all correct guesses are made by second method as the probability of making all correct guesses is $\left(\frac{1}{4}\right)^4 = \frac{1}{16} = 0.06$ it is easy to see that if 10 chapaties are given by second method and if all correct guesses are made, then we can reject the null hypothesis at 0.05 level of significance. Not to unduly influence the taster in making guesses, we should also present the chapaties in a random order rather than systematically presenting them for tasting.

The above discussed chapati tasting experiment brings home the following salient features of experimentation:

- All the extraneous variations in the data should be eliminated or controlled excepting the variations due to the treatments under study. One should not artificially provide circumstances for one treatment to show better results than others.
- Far a given size of the experiment, though the experiment can be done in many ways, even the best results may not turn out to be significant with some designs, while some other design can detect the treatment differences. Thus there is an imperative need the choose the right type of design, before the commencement of the experiment, lest the results may be useless.
- If for some specific reasons related to the nature .of the experiment, a particular method has to be used in experimentation, then adequate number of replications of each treatment have to be provided in order to get valid inferences.
- The treatments have to be randomly allocated to the experimental units.

The terminologies often used in planning and designing of experiments are listed below.

Treatment

Treatment refers to controllable quantitative or qualitative factors imposed at a certain level by the experimenter. For an agronomist several fertilizer concentrations applied to a particular crop or a variety of crop is a treatment. Similarly, an animal scientist looks upon several concentrations of a drug given to animal species as a treatment. In agribusiness we may look upon impact of advertising strategy on sales a treatment. To an agricultural engineer, different levels of irrigation may constitute a treatment.

Experimental Unit

An experimental unit is an entity that receives a treatment e.g., for an agronomist or horticulturist it may be a plot of a land or batch of seed, for an animal scientist it may be a group of pigs or sheep, for a scientist engaged in forestry research it may be different tree species occurring in an area, and for an agricultural engineer it may be manufactured item. Thus, an experimental unit maybe looked upon as a small subdivision of the experimental material, which receives the treatment.

Experimental Error

Differences in yields arising out of experimental units treated alike are called Experimental Error.

Controllable conditions in an experiment or experimental variable are terms as a factor. For example, a fertilizer, a new feed ration, and a fungicide are all considered as factors. Factors may be qualitative or quantitative and may take a finite number of values or type. Quantitative factors are those described by numerical values on some scale. The rates of application of fertilizer, the quantity of seed sown are examples of quantitative factors. Qualitative factors are those factors that can be distinguished from each other, but not on numerical scale e.g., type of protein in a diet, sex of an animal, genetic make up of plant etc. While choosing factors for any experiment researcher should ask the following questions, like What treatments in the experiment should be related directly to the objectives of the study? Does the experimental technique adopted require the use of additional factors? Can the experimental unit be divided naturally into groups such that the main treatment effects are different for the different groups? What additional factors should one include in the experiment to interact with the main factors and shed light on the factors of direct interest? How desirable is it to deliberately choose experimental units of different types?

Basic Principles of Design of Experiments

Given a set of treatments which can provide information regarding the objective of an experiment, a design for the experiment, defines the size and number of experimental units, the manner in which the treatments are allotted to the units and also appropriate type and grouping of the experimental units. These requirements of a design ensure validity, interpretability and accuracy of the results obtainable from an analysis of the observations.

These purposes are served by the principles of:

- Randomization
- Replication
- Local (Error) control

Randomization

After the treatments and the experimental units are decided the treatments are allotted to the experimental units at random to avoid any type of personal or subjective bias, which may be conscious or unconscious. This ensures validity of the results. It helps to have an objective comparison among the treatments. It also ensures independence of the observations, which is necessary for drawing valid inference from the observations by applying appropriate statistical techniques.

Depending on the nature of the experiment and the experimental units, there are various experimental designs and each design has its own way of randomization. Various speakers while discussing specific designs in the lectures to follow shall discuss the procedure of random allocation separately.

Replication

If a treatment is allotted to r experimental units in an experiment, it is said to be replicated r times. If in a design each of the treatments is replicated r times, the design is said to have r replications. Replication is necessary to

- Provide an estimate of the error variance which is a function of the differences among observations from experimental units under identical treatments.
- Increase the accuracy of estimates of the treatment effects.

Though, more the number of replications the better it is, so far as precision of estimates is concerned, it cannot be increased infinitely as it increases the cost of experimentation. Moreover, due to limited availability of experimental resources too many replications cannot be taken.

The number of replications is, therefore, decided keeping in view the permissible expenditure and the required degree of precision. Sensitivity of statistical methods for drawing inference also depends on the number of replications. Sometimes this criterion is used to decide the number of replications in specific experiments.

Error variance provides a measure of precision of an experiment, the less the error variance the more precision. Once a measure of error variance is available for a set of experimental units, the number of replications needed for a desired level of sensitivity can be obtained as below.

Given a set of treatments an experimenter may not be interested to know if two treatment differ in their effects by less than a certain quantity, say, d. In other words, he wants an experiment that should be able to differentiate two treatments when they differ by d or more.

The significance of the difference between two treatments is tested by t-test where

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{2s^2/r}},$$

Here, $\bar{y}_i$, and $\bar{y}_j$ are the arithmetic means of two treatment effects each based on r replications, s^2 is measure of error variation.

Given a difference d, between two treatment effects such that any difference greater than d should be brought out as significant by using a design with r replications, the following equation provides a solution of r.

$$t = \frac{|d\ |}{\sqrt{2s^2/r}},$$

$$r = \frac{t_0^2}{d^2} x 2s^2 \qquad \dots(1)$$

where t_0 is the critical value of the t-distribution at the desired level of significance, that is, the value of t at 5 or 1 per cent level of significance read from the t-table. If s^2 is known or based on a very large number of observations, made available from some pilot pre-experiment investigation, then t is taken as the normal variate. If s^2 is estimated with n degree of freedom (d.f.) then t_0 corresponds to n d.f.

When the number of replication is r or more as obtained above, then all differences greater than d are expected to be brought out as significant by an experiment when it is conducted on a set of experimental units which has variability of the order of s^2. For example, in an experiment on wheat crop conducted in a seed farm in Bhopal, to study the effect of application of nitrogen and phosphorous on yield a randomized block design with three replications was adopted. There were 11 treatments two of which were (i) 60 Kg/ha of nitrogen (ii) 120 Kg/ha of nitrogen. The average yield figures for these two application of the fertilizer were 1438 and 1592 Kg/ha respectively and it is required that differences of the order of 150 Kg/ha should be brought out significant. The error mean square (s^2) was 12134.88. Assuming that the experimental error will be of the same order in future experiments and t_0 is of the order of 2.00, which is likely as the error d.f. is likely to be more than 30 as there are 11 treatments; Substituting in (1), we get:

$$r = \frac{2t_0^2 s^2}{d^2} = \frac{2x2^2 x2134.88}{150^2} = 4\,(\text{approx.})$$

Thus, an experiment with 4 replications is likely to bring out differences of the order of 150 Kg/ha as significant.

Another criterion for determining r is to take a number of replications which ensures at least 10 d.f. for the estimate of error variance in the analysis of variance of the design concerned since the sensitivity of the experiment will be very much low as the F test (which is used to draw inference in such experiments) is very much unstable below 10 d.f.

Local Control

The consideration in regard to the choice of number of replications ensure reduction of standard error of the estimates of the treatment effect because the standard error of the estimate of a treatment effect is $\sqrt{s^2 / r}$, but it cannot reduce the error variance itself. It is, however, possible to devise methods for reducing the error variance. Such measures are called *error control* or local control. One such measure is to make the experimental units homogenous. Another method is to form the units into several homogenous groups, usually called blocks, allowing variation between the groups.

A considerable amount of research work has been done to divide the treatments into suitable groups of experimental units so that the treatment effect can be estimated more precisely Extensive use of combinatorial mathematics has been made for formation of such group treatments. This grouping of experiment units into different groups has led to the development of various designs useful to the experimenter. We now briefly describe the various term used in designing of an experiment

Blocking

It refers to methodologies that form the units into homogeneous or pre-experimental subject-similarity groups. It is a method to reduce the effect of variation in the experimental material on the Error of Treatment of Comparisons. For example, animal scientist may decide to group animals on age, sex, breed or some other factors that he may believe has an influence on characteristic being measured. Effective blocking removes considerable measure of variation nom the experimental error. The selection of source of variability to be used as basis of blocking, block size, block shape and orientation are crucial for blocking. The blocking factor is introduced in the experiment to increase the power of design to detect treatment effects.

The importance of good designing is inseparable from good research (results). The following examples point out the necessity for a good design that will yield good research. First, a nutrition specialist in developing country is interested in determining whether mother's milk is better than powdered milk for children under age one. The nutritionist has compared the growth of children in village A, who are all on mother's milk against the children in village B, who use powdered milk. Obviously, such a comparison ignores the health of the mothers, the sanitary-conditions of the villages, and other factors that may have contributed to the differences observed without any connection to the advantages of mother's milk or the powdered milk on the children. A proper design would require that both mother's milk and the powdered milk be alternatively used in both villages, or some other methodology to make certain that the differences observed are attributable to the type of milk consumed and not to some uncontrollable factor. Second, a crop scientist who is comparing 2 varieties of maize, for instance, would not assign one variety to a location where such factors as sun, shade, unidirectional fertility gradient, and uneven distribution of water would either favor or handicap it over the other. If such a design were to be adopted, the researcher would have difficulty in determining whether the apparent difference in yield was due to variety differences or resulted from such factors as sun, shade, soil fertility of the field, or the distribution of water. These two examples illustrate the type of poorly designed experiments that are to be avoided.

Analysis of Variance

Analysis of Variance (ANOVA) is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned and is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the *response variable is continuous* in nature, whereas the *predictor variables are categorical*. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In a particular case, where two population means are being compared, ANOVA is equivalent to the independent two-sample *t*-test.

The fixed-effects model of ANOVA applies to situations in which the experimenter applies several treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole. In it factors are fixed and are attributable to a finite set of levels of factor eg. Sex, year, variety, fertilizer etc.

Consider for example a clinical trial where three drugs are administered on a group of men and women some of whom are married and some are unmarried. The three classifications of sex, drug and marital status that identify the source of each datum are known as factors. The individual classification of each factor is known as levels of the factors. Thus, in this example there are 3 levels of factor drug, 2 levels of factor sex and 2 levels of marital status. Here all the effects are fixed. Random effects models are used when the treatments are not fixed. This occurs when the various treatments (also known as factor levels) are sampled from a larger population.

When factors are random, these are generally attributable to infinite set of levels of a factor of which a random sample are deemed to occur *eg.* research stations, clinics in Delhi, sire, etc. Suppose new inject-able insulin is to be tested using 15 different clinics of Delhi state. It is reasonable to assume that these clinics are random sample from a population of clinics from Delhi. It describe the situations where both fixed and random effects are present.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of *t*-tests: a comparison of differences of means across more than two groups.

The ANOVA is valid under certain assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2.
- Effects are additive in nature.

The ANOVA is performed as one-way, two-way, three-way, etc. ANOVA when the number of factors is one, two or three respectively. In general if the number of factors is more, it is termed as multi-way ANOVA.

BASIC EXPERIMENTAL DESIGNS

1. Introduction

In this chapter, three basic designs viz., Completely randomized design (CRD), Randomized Complete Block Design (RCBD) and Latin Square Design (LSD) are explained in detail.

2. Completely Randomized Design

Designs are usually characterized by the nature of grouping of experimental units and the procedure of random allocation of treatments to the experimental units. In a completely randomized design the units are taken in a single group. As far as possible the units forming the group are homogeneous. This is a design in which only randomization and replication are used. There is no use of local control here.

Let there be *v* treatments in an experiment and *n* homogeneous experimental units. Let the i^{th} treatment be replicated r_i times $(i = 1,2,..., v)$ such that $\sum_{i=1}^{v} r_i = n$. The treatments are allotted at random to the units.

Normally the number of replications for different treatments should be equal as it ensures equal precision of estimates of the treatment effects. The actual number of replications is, however, determined by the availability of experimental resources and the requirement of precision and sensitivity of comparisons. If the experimental material for some treatments is available in limited quantities, the numbers of their replication are reduced. If the estimates of certain treatment effects are required with more precision, the numbers of their replication are increased.

Randomization

There are several methods of random allocation of treatments to the experimental units. The *v* treatments are first numbered in any order from *1* to *v*. The *n* experimental units are also numbered suitably. One of the methods uses the random number tables. Any page of a random number table is taken. If *v* is a one-digit number, then the table is consulted digit by digit. If *v* is a two-digit number, then two-digit random numbers are consulted. All numbers greater than *v* including zero are ignored.

Let the first number chosen be n_1; then the treatment numbered n_1 is allotted to the first unit. If the second number is n_2 which may or may not be equal to n_1 then the treatment numbered n_2 is allotted to the second unit. This procedure is continued. When the i^{th} treatment number has occurred r_i times, $(i = 1,2,..., v)$ this treatment is ignored subsequently. This process terminates when all the units are exhausted.

One drawback of the above procedure is that sometimes a very large number of random numbers may have to be ignored because they are greater than *v*. It may even happen that the random number table is exhausted before the allocation is complete. To avoid this difficulty the following procedure is adopted. We have described the procedure by taking *v* to be a two-digit number.

Let *P* be the highest two-digit number divisible by *v*. Then all numbers greater than *P* and zero are ignored. If a selected random number is less than *v*, then it is used as such. If it is greater than or equal to *v*, then it is divided by *v* and the remainder is taken to the random number. When a number is completely divisible by *v*, then the random number is *v*. If *v* is an *n*-digit number, then *P* is taken to be the highest *n*-digit number divisible by *v*. The rest of the procedure is the same as above.

Analysis

This design provides a one-way classified data according to levels of a single factor. For its analysis the following model is taken:

$$y_{ij} = \mu + t_i + e_{ij}\ , \qquad i = 1,\cdots,v;\ j = 1,\cdots r_i\,,$$

where y_{ij} is the random variable corresponding to the observation y_{ij} obtained from the j^{th} replicate of the i^{th} treatment, μ is the general mean, t_i is the fixed effect of the i^{th} treatment and e_{ij} is the error component which is a random variable assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Let $\sum_j y_{ij} = T_i \quad (i = 1,2,...,v)$ be the total of observations from i^{th} treatment. Let further $\sum_i T_i = G.$ Correction factor (*C.F.*) $= G^2/n.$

Sum of squares due to treatments $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$

Total sum of squares = $\sum_{i=1}^{v} \sum_{j=1}^{r_i} y_{ij}^2 - C.F.$

Analysis of Variance

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Treatments	$v - 1$	*SST* $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$	*MST = SST / (v - 1)*	*MST/MSE*
Error	$n - v$	*SSE = by subtraction*	*MSE = SSE / (n - v)*	
Total	$n - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test where F is the ratio *MST / MSE* with *(v - 1)* and *(n - v)* degrees of freedom.

3. Randomized Complete Block Design

It has been seen that when the experimental units are homogeneous then a CRD should be adopted. In any experiment, however, besides treatments the experimental material is a major source of variability in the data. When experiments require a large number of experimental units, the experimental units may not be homogeneous, and in such situations CRD can not be recommended. When the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or blocks). The principle of first forming homogeneous groups of the experimental units and then allotting at random each treatment once in each group is known as local control. This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks, is smaller because of homogeneous blocks. This type of allocation makes it possible to eliminate from error variance a portion of variation attributable to block differences. If, however, variation between the blocks is not significantly large, this type of grouping of the units does not lead to any advantage; rather some degrees of freedom of the error variance is lost without any consequent decrease in the error variance. In such situations it is not desirable to adopt randomized complete block designs in preference to completely randomized designs.

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group then such an arrangement is called a ***randomized complete block design.***

Suppose the experimenter wants to study *v* treatments. Each of the treatments is replicated *r* times (the number of blocks) in the design. The total number of experimental units is, therefore, *vr*. These units are arranged into *r* groups of size *v* each. The error control measure in this design consists of making the units in each of these groups homogeneous.

The number of blocks in the design is the same as the number of replications. The *v* treatments are allotted at random to the *v* plots in each block. This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block designs. The availability of resources and considerations of cost and precision determine actual number of replications in the design.

Analysis

The data collected from experiments with randomized complete block designs form a two-way classification, that is, classified according to the levels of two factors, *viz.,* blocks and treatments. There are *vr* cells in the two-way table with one observation in each cell. The data are orthogonal and therefore the design is called an *orthogonal design.* We take the following model:

$$y_{ij} = \mu + t_i + b_j + e_{ij} \ , \qquad \begin{pmatrix} i = 1,2,...,v; \\ j = 1,2,...,r \end{pmatrix},$$

where y_{ij} denotes the observation from i^{th} treatment in j^{th} block. The fixed effects μ, t_i, b_j denote respectively the general mean, effect of the i^{th} treatment and effect of the j^{th} block. The random variable e_{ij} is the error component associated with y_{ij}. These are assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Following the method of analysis of variance for finding sums of squares due to blocks, treatments and error for the two-way classification, the different sums of squares are obtained as follows: Let $\sum_j y_{ij} = T_i \ (i = 1,2,...,v)$ = total of observations from i^{th} treatment and $\sum_j y_{ij} = B_j$ $j = 1,\cdots,r$ = total of observations from j^{th} block. These are the marginal totals of the two-way data table. Let further, $\sum_i T_i = \sum_j B_j = G.$

Correction factor ($C.F.$) = G^2/rv, Sum of squares due to treatments $= \sum_i \frac{T_i^2}{r} - C.F.$,

Sum of squares due to blocks $= \sum_j \frac{B_j^2}{v} - C.F.$, Total sum of squares $= \sum_{ij} y_{ij}^2 - C.F.$

Analysis of Variance

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Blocks	$r - 1$	$SSB = \sum_j \frac{B_j^2}{v} - C.F.$	$MSB = SSB / (r - 1)$	MSB/MSE
Treatments	$v - 1$	$SST = \sum_i \frac{T_i^2}{r} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$(r - 1)(v - 1)$	SSE = *by subtraction*	$MSE = SSE / (v - 1)(r - 1)$	
Total	$vr - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test, where F is the ratio *MST / MSE* with *(v - 1)* and *(v - 1)(r - 1)* degrees of freedom. We may then be interested to either compare the treatments in pairs or evaluate special contrasts depending upon the objectives of the experiment. This is done as follows:

The critical difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $C.D. = t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE/r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's t at the level of significance α and degree of freedom *(v - 1)(r - 1)*. If the difference of any two-treatment means is greater than the C.D. value, the corresponding treatment effects are significantly different.

4. Latin Square Design

Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. These designs require that the number of replications equal the number of ***treatments*** or ***varieties***.

Definition 1. A Latin square arrangement is an arrangement of v symbols in v^2 cells arranged in v rows and v columns, such that every symbol occurs precisely once in each row and precisely once in each column. The term v is known as the **order** of the Latin square.

If the symbols are taken as *A, B, C, D,* a Latin square arrangement of order 4 is as follows:

$$\begin{array}{cccc} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{array}$$

A Latin square is said to be in the ***standard form*** if the symbols in the first row and first column are in natural order, and it is said to be in the ***semi-standard form*** if the symbols of the first row are in natural order. Some authors denote both of these concepts by the term ***standard form***. However, there is a need to distinguish between these two concepts. The standard form is used for randomizing the Latin-square designs, and the semi-standard form is needed for studying the properties of the orthogonal Latin squares.

Definition 2. If in two Latin squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, the two Latin squares are said to be **orthogonal**. If the symbols of one Latin square are denoted by Latin letters and the symbols of the other are denoted by Greek letters, the pair of orthogonal Latin squares is also called a **graeco-latin square**.

Definition 3. If in a set of Latin squares every pair is orthogonal, the set is called a set of **mutually orthogonal latin squares (MOLS)**. It is also called a **hypergraeco latin square.**

The following is an example of graeco latin square:

$$\begin{array}{cccc} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{array} \qquad \begin{array}{cccc} \alpha & \gamma & \delta & \beta \\ \beta & \delta & \gamma & \alpha \\ \gamma & \alpha & \beta & \delta \\ \delta & \beta & \alpha & \gamma \end{array} \qquad \begin{array}{cccc} A\alpha & B\gamma & C\delta & D\beta \\ B\beta & A\delta & D\gamma & C\alpha \\ C\gamma & D\alpha & A\beta & B\delta \\ D\delta & C\beta & B\alpha & A\gamma \end{array}$$

We can verify that in the above arrangement every pair of ordered Latin and Greek symbols occurs exactly once, and hence the two latin squares under consideration constitute a graecolatin square.

It is well known that the maximum number of MOLS possible of order v is v - 1. A set of v - 1 MOLS is known as a complete set of MOLS. Complete sets of MOLS of order v exist when v is a ***prime or prime power.***

Randomization

According to the definition of a Latin square design, treatments can be allocated to the v^2 experimental units (may be animal or plots) in a number of ways. There are, therefore, a number of Latin squares of a given order. The purpose of randomization is to select one of these squares at random. The following is one of the methods of random selection of Latin squares.

Let a $v \times v$ Latin square arrangement be first written by denoting treatments by Latin letters *A, B, C, etc.* or by numbers *1, 2, 3, etc.* Such arrangements are readily available in the ***Tables for Statisticians and Biometricians*** (Fisher and Yates, 1974). One of these squares of any order can be written systematically as shown below for a 5×5 Latin square:

$$\begin{array}{ccccc} A & B & C & D & E \\ B & C & D & E & A \\ C & D & E & A & B \\ D & E & A & B & C \\ E & A & B & C & D \end{array}$$

For the purpose of randomization rows and columns of the Latin square are rearranged randomly. There is no randomization possible within the rows and/or columns. For example, the following is a row randomized square of the above 5×5 Latin square;

$$\begin{array}{ccccc} A & B & C & D & E \\ B & C & D & E & A \\ E & A & B & C & D \\ D & E & A & B & C \\ C & D & E & A & B \end{array}$$

Next, the columns of the above row randomized square have been rearranged randomly to give the following random square:

$$\begin{array}{ccccc} E & B & C & A & D \\ A & C & D & B & E \\ D & A & B & E & C \\ C & E & A & D & B \\ B & D & E & C & A \end{array}$$

As a result of row and column randomization, but not the randomization of the individual units, the whole arrangement remains a Latin square.

Analysis

In Latin square designs there are three factors. These are the factors *P*, *Q*, and treatments. The data collected from this design are, therefore, analyzed as a three-way classified data. Actually, there should have been v^3 observations as there are three factors each at *v* levels. But because of the particular allocation of treatments to the cells, there is only one observation per cell instead of *v* in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained. Accordingly, we take the model

$$Y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} denotes the observation in the i^{th} row, j^{th} column and under the s^{th} treatment; $\mu, r_i, c_j, t_s (i, j, s = 1,2,...,v)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component, assumed to be independently and normally distributed with zero mean and a constant variance, σ^2.

The analysis is conducted by following a similar procedure as described for the analysis of two-way classified data. The different sums of squares are obtained as below: Let the data be arranged first in a *row* × *column* table such that y_{ij} denotes the observation of (*i*, *j)*th cell of table.

Let $R_i = \sum_j y_{ij} = i^{th}\ rowtotal(i = 1,2,...,v), C_j = \sum_i y_{ij} = j^{th}\ columntotal(j = 1,2,...,v),\quad T_s =$ sum of those observations which come from s^{th} treatment *(s= 1,2,…,v),*

$G = \sum_i R_i = grand total.$ Correction factor, $C.F. = \frac{G^2}{v^2}$. Treatment sum of squares = $\sum_s \frac{T_s^2}{v} - C.F.$, Row sum of squares = $\sum_i \frac{R_i^2}{v} - C.F.$, Column sum of squares = $\sum_j \frac{C_j^2}{v} - C.F.$

Analysis of Variance of $v \times v$ Latin Square Design

Sources of Variation	D.F.	S.S.	M.S.	F
Rows	v -1	$\sum_i \frac{R_i^2}{v} - C.F.$		
Columns	v - 1	$\sum_j \frac{C_j^2}{v} - C.F.$		
Treatments	v - 1	$\sum_s \frac{T_s^2}{v} - C.F.$	s_t^2	s_t^2 / s_e^2
Error	(v - 1)(v - 2)	By subtraction	s_e^2	
Total	v^2-1	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis of equal treatment effects is tested by *F*-test, where *F* is the ratio of treatment mean squares to error mean squares. If *F* is not significant, treatment effects do not differ significantly among themselves. If *F* is significant, further studies to test the significance of any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

5. Contrasts Analysis

The main technique adopted for the analysis and interpretation of the data collected from an experiment is the analysis of variance technique that essentially consists of partitioning the total variation in an experiment into components ascribable to different sources of variation due to the controlled factors and error. Analysis of variance clearly indicates a difference among the treatment means. The objective of an experiment is often much more specific than merely determining whether or not all of the treatments give rise to similar responses. For examples, a chemical experiment might be run primarily to determine whether or not the yield of the chemical process increases as the amount of the catalyst is increased. A medical experimenter might be concerned with the efficacy of each of several new drugs as compared to a standard drug. A nutrition experiment may be run to compare high fiber diets with low fiber diets. A plant breeder may be interested in comparing exotic collections with indigenous cultivars. An agronomist may be interested in comparing the effects of biofertilisers and chemical fertilisers. A water technologist may be interested in studying the effect of nitrogen with Farm Yard Manure over the nitrogen levels without farm yard manure in presence of irrigation.

2.1 Contrasts

Let $y_1, y_2, ..., y_n$ denote n observations or any other quantities. The linear function $C = \sum_{i=1}^{n} l_i y_i$, where l_i's are given number such that $\sum_{i=1}^{n} l_i = 0$, is called a *contrast* of $y_i' s$. Let $y_1, y_2, ..., y_n$ be independent random variables with a common mean μ and variance σ^2. The expected value of the random variable C is zero and its variance is $\sigma^2 \sum_{i-1}^{n} l_i^2$. In what follows we shall not distinguish between a contrast and its corresponding random variable.

Sum of squares (s.s.) of contrasts. The sum of squares due to the contrast C is defined as $C^2 / \sigma^{-2} Var(C) = C^2 / \left(\sum_{i=1}^{n} l_i^2 \right)$. Here σ^2 is unknown and is replaced by its unbiased estimate, *i.e. mean square error*. It is known that this square has a $\sigma^2 \chi^2$ distribution with one degree of freedom when the $y_i' s$ are normally distributed. Thus the sum of squares due to two or more contrasts has also a $\sigma^2 \chi^2$ distribution if the contrasts are independent. Multiplication of any contrast by a constant does not change the contrast. The sum of squares due to a contrast as defined above is not evidently changed by such multiplication.

Orthogonal contrasts. Two contrasts, $C_1 = \sum_{i=1}^{n} l_i y_i$ and $C_2 = \sum_{i=1}^{n} l_i y_i$ are said to be orthogonal if and only if $\sum_{i=1}^{n} l_i m_i = 0$. This condition ensures that the covariance between C_1 and C_2 is zero.

When there are more than two contrasts, they are said to be mutually orthogonal if they are orthogonal pair wise. For example, with four observations y_1, y_2, y_3, y_4, we may write the following three mutually orthogonal contrasts:

(i) $y_1 + y_2 - y_3 - y_4$

(ii) $y_1 - y_2 - y_3 + y_4$

(iii) $y_1 - y_2 + y_3 - y_4$

The sum of squares due to a set of mutually orthogonal contrasts has a $\sigma^2 \chi^2$ distribution with as many degrees of freedom as the number of contrasts in the set.

ANALYSIS OF COVARIANCE

Introduction

The meaning of ANVOVA is Analysis of Covariance. It is a general linear model with one continuous outcome variable (quantitative) and one or more factor variables (qualitative). ANCOVA is a merger of ANOVA and regression for continuous variables. ANCOVA tests whether certain factors have an effect on the outcome variable after removing the variance for which quantitative predictors (covariates) account. The inclusion of covariates can increase statistical power because it accounts for some of the variability.

It is well known that in designed experiments the ability to detect existing differences among treatments increases as the size of the experimental error decreases, a good experiment attempts to incorporate all possible means of minimizing the experimental error. Besides proper experimentation, a proper data analysis also helps in controlling experimental error. In situations where blocking alone may not be able to achieve adequate control of experimental error, proper choice of data analysis may help a great deal. By measuring one or more *covariates* - the characters whose functional relationships to the character of primary interest are known - the Analysis of Covariance (ANCOVA) can reduce the variability among experimental units by adjusting their values to a common value of the covariates. For example, in an animal feeding trial, the initial body weight of the animals usually differs. Using this initial body weight as a covariate, the final weights recorded after the animals have been subjected to various physiological feeds (treatments) can be adjusted to the values that would have been obtained had there been no variation in the initial body weights of the animals at the start of the experiment. An another example, in a field experiment where rodents have (partially) damaged some of the plots, covariance analysis with rodent damage as a covariate could be useful in adjusting plot yields to the levels that they should have been had there been no rodent damage in any plot.

ANCOVA requires measurement of the character of primary interest plus the measurement of one or more variables known as *covariates*. It also requires that the functional relationship of the covariates with the character of primary interest is known beforehand. Generally a linear relationship is assumed, though other type of relationships could also be assumed.

Consider the case of a variety trial in which weed incidence is used as a covariate. With a known functional relationship between weed incidence and grain yield, the character of primary interest, the covariance analysis can adjust grain yield in each plot to a common level of weed incidence. With this adjustment, the variation in yield due to weed incidence is quantified and effectively separated from that due to varietal difference.

ANCOVA can be applied to any number of covariates and to any type of functional relationship between variables *viz.* quadratic, inverse polynomial, etc. Here we illustrate the use of covariance analysis with the help of a single covariate that is linearly related with the character of primary interest. It is expected that this simplification shall not unduly reduce the applicability of the technique, as a single covariate that is linearly related with the primary variable is adequate for most of the experimental situations in agricultural research.

Uses of Covariance Analysis in Agricultural Research

There are several important uses of covariance analysis in agricultural research. Some of the most important ones are:

1. To control experimental error and to adjust treatment means.
2. To aid in the interpretation of experimental results.
3. To estimate missing data.

Error Control and Adjustment of Treatment Means

It is now well realized that the size of experimental error is closely related to the variability between experimental units. It is also known that proper blocking can reduce experimental error by maximizing the differences between the blocks and thus minimizing differences within blocks. Blocking, however, can not cope with certain types of variability such as spotty soil heterogeneity and unpredictable insect incidence. In both instances, heterogeneity between experimental plots does not follow a definite pattern, which causes difficulty in getting maximum differences between blocks. Indeed, blocking is ineffective in the case of nonuniform insect incidences because blocking must be done before the incidence occurs. Furthermore, even though it is true that a researcher may have some information on the probable path or direction of insect movement, unless the direction of insect movement coincides with the soil fertility gradient, the choice of whether soil heterogeneity or insect incidence should be the criterion for blocking is difficult. The choice is especially difficult if both sources of variation have about the same importance.

Use of covariance analysis should be considered in experiments in which blocking couldn't adequately reduce the experimental error. By measuring an additional variable (*e.g.,* covariate X) that is known to be linearly related to the primary variable Y, the source of variation associated with the covariate can be deducted from experimental error. This adjusts the primary variable Y linearly upward or downward, depending on the relative size of its respective covariate. The adjustment accomplishes two important improvements:

1. The treatment mean is adjusted to a value that it would have had; had there been no differences in the values of the covariate.
2. The experimental error is reduced and the precision for comparing treatment means is increased.

Although blocking and covariance techniques are both used to reduce experimental error, the differences between the two techniques are such that they are usually not interchangeable. The ANCOVA can be used only when the covariate representing the heterogeneity among the experimental units can be measured quantitatively. However, that is not a necessary condition for blocking. In addition, because blocking is done before the start of the experiment, it can be used only to cope with sources of variation that are known or predictable. ANCOVA, on the other hand, can take care of unexpected sources of variation that occur during the experiment. Thus, ANCOVA is useful, as a supplementary procedure to take care of sources of variation that cannot be accounted for by blocking.

When covariance analysis is used for error control and adjustment of treatment means, the covariate must not be affected by the treatments being tested. Otherwise, the adjustment removes both the variation due to experimental error and that due to treatment effects. A good example of covariates that are free of treatment effects are those that are measured before the treatments are applied, such as soil analysis and residual effects of treatments applied in the past experiments. In other cases, care must be exercised to ensure that the covariates defined are not affected by the treatments being tested. This technique can be illustrated through the following example:

Example 1: A trial was designed to evaluate 15 rice varieties grown in soil with a toxic level of iron. The experiment was in a RCB design with three replications. Guard rows of a susceptible check variety were planted on two sides of each experimental plot. Scores for tolerance for iron toxicity were collected from each experimental plot as well as from guard rows. For each experimental plot, the score of susceptible check (averaged over two guard rows) constitutes the value of the covariate for that plot. Data on the tolerance score of each variety (Y variable) and on the score of the corresponding susceptible check (X variable) are shown below:

Scores of tolerance for iron toxicity (Y) of 15 rice varieties and those of the corresponding guard rows of a susceptible check variety (X) in a RCB trial

Variety Number	Replication-I		Replication-II		Replication-III	
	X	Y	X	Y	X	Y
1.	15	22	16	13	16	14
2.	16	14	15	23	15	23
3.	15	24	15	24	15	23
4.	16	13	15	23	15	23
5.	17	17	17	16	16	16
6.	16	14	15	23	15	23
7.	16	13	15	23	16	13
8.	16	16	17	17	16	16
9.	17	14	15	23	15	24
10.	17	17	17	17	15	26
11.	16	15	15	24	15	25
12.	16	15	15	23	15	23
13.	15	24	15	24	16	15
14.	15	25	15	24	15	23
15.	15	24	15	25	16	16

The usual analysis of variance without using the covariate (X variable) is as follows:

Source	DF	SS	Mean Square	F Value	Pr > F
Replication	2	104.0444	52.0222	2.85	0.0745
Treatment	14	265.9111	18.9937	1.04	0.4448
Error	28	510.6222	18.2365		
Total	**44**	**880.5778**			

R-Square	C.V.	Root MSE	Y - Mean
0.4201	21.5436	4.2704	19.82222

Using the covariate, the analysis is the following:

Source	DF	S.S.	M.S.	F-Value	Pr > F
Replication	2	22.4802	11.2402	2.71	0.0844
Treatment	14	152.5606	10.8972	2.63	0.0151
Covariate X	1	398.7516	398.7516	96.24	0.0001
Error	27	111.8707	4.1434		

R-Square	C.V.	Root MSE	Y Mean
0.8730	10.2689	2.0355	19.8222

It is interesting to note that the use of a covariate has resulted into a considerable reduction in the error mean square and hence the CV has also reduced drastically. This has helped in catching the small differences among the treatment effects as significant. This was not possible when the covariate was not used. The covariance analysis will thus result into a more precise comparison of treatment effects.

The probability of significance of pairwise comparisons among the least square estimates of the treatment effects are given below:

```
                          Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
i/j     1       2       3       4       5       6       7       8       9
1       .     0.3370  0.0666  0.4431  0.0019  0.3370  1.0000  0.0252  0.0232
2     0.3370    .     0.3370  0.8425  0.0237  1.0000  0.3370  0.1834  0.1697
3     0.0666  0.3370    .     0.2497  0.1620  0.3370  0.0666  0.6757  0.6751
4     0.4431  0.8425  0.2497    .     0.0157  0.8425  0.4431  0.1320  0.1191
5     0.0019  0.0237  0.1620  0.0157    .     0.0237  0.0019  0.2361  0.2493
6     0.3370  1.0000  0.3370  0.8425  0.0237    .     0.3370  0.1834  0.1697
7     1.0000  0.3370  0.0666  0.4431  0.0019  0.3370    .     0.0252  0.0232
8     0.0252  0.1834  0.6757  0.1320  0.2361  0.1834  0.0252    .     0.9727
9     0.0232  0.1697  0.6751  0.1191  0.2493  0.1697  0.0232  0.9727    .
10    0.0001  0.0019  0.0237  0.0012  0.3370  0.0019  0.0001  0.0361  0.0385
11    0.0874  0.4294  0.8575  0.3249  0.1046  0.4294  0.0874  0.5445  0.5439
12    0.2497  0.8425  0.4431  0.6915  0.0351  0.8425  0.2497  0.2493  0.2361
13    0.1270  0.5524  0.7066  0.4294  0.0739  0.5524  0.1270  0.4298  0.4229
14    0.0446  0.2497  0.8425  0.1803  0.2158  0.2497  0.0446  0.8096  0.8204
15    0.0589  0.3249  0.9860  0.2393  0.1452  0.3249  0.0589  0.6736  0.6809
```

```
 Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
   i/j      10      11      12        13        14        15
    1     0.0001  0.0874  0.2497    0.1270    0.0446    0.0589
    2     0.0019  0.4294  0.8425    0.5524    0.2497    0.3249
    3     0.0237  0.8575  0.4431    0.7066    0.8425    0.9860
    4     0.0012  0.3249  0.6915    0.4294    0.1803    0.2393
    5     0.3370  0.1046  0.0351    0.0739    0.2158    0.1452
    6     0.0019  0.4294  0.8425    0.5524    0.2497    0.3249
    7     0.0001  0.0874  0.2497    0.1270    0.0446    0.0589
    8     0.0361  0.5445  0.2493    0.4298    0.8096    0.6736
    9     0.0385  0.5439  0.2361    0.4229    0.8204    0.6809
   10       .     0.0124  0.0031    0.0079    0.0351    0.0191
   11     0.0124    .     0.5524    0.8425    0.7066    0.8425
   12     0.0031  0.5524    .       0.6915    0.3370    0.4294
   13     0.0079  0.8425  0.6915      .       0.5671    0.6915
   14     0.0351  0.7066  0.3370    0.5671      .       0.8575
   15     0.0191  0.8425  0.4294    0.6915    0.8575      .
```

ANALYSIS OF REPEATED MEASURES DATA

1. Introduction

The term “repeated measures” refers broadly to the data in which the response of each experimental unit or subject is observed on multiple occasions or under multiple conditions. Thus repeated measurements refer to the situation in which multiple measurements of the response variable are obtained, over several time periods, from each experimental unit, such as an animal. Usually, the responses are taken over time, as in growth of animal weights are measured weekly/monthly production of fruit over the years from the same tree. Repeated measurement data are obtained in animal science, horticulture, clinical trials, medical science, physiological, psychological experiments, etc.

Repeated measures experiments are a type of factorial experiment, with group and time as the two factors. They have been used commonly in animal, plant, and human research for several decades, but only in recent years statistical and computing methodologies been available to analyze them effectively and efficiently. The objectives of repeated measures data analysis are to examine and compare response trends over time. This can involve comparisons of groups at specific times, or averaged over time. It also can involve comparisons of times within a group. These are objectives common to any factorial experiment. The important feature of repeated measures experiments that requires special attention in data analysis is the correlation pattern among the responses on the same individual (animal) over time.

2. Methods for Analyzing Repeated Measures

Responses measured on the same animal are correlated because they contain a common contribution from the animal. Moreover, measures on the same animal close in time tend to be more highly correlated than measures far apart in time. Also, variances of repeated measures often change with time. These potential patterns of correlation and variation may combine to produce a complicated covariance structure of repeated measures. Special methods of statistical analysis are needed for repeated measures data because of the covariance structure. Standard regression and analysis of variance methods may produce invalid results because they require mathematical assumptions that do not hold with repeated measures data. In repeated measures analysis of variance, the effects of interest are

i) between-subject effects such as GROUP
ii) within-subject effects such as TIME
iii) interactions between the two types of effects such as GROUP*TIME.

There are several statistical methods used for analyzing repeated measures data. Here we give from basic to sophisticated methods for the analysis of repeated measure data using SAS software. These include:

i) Separate analyses at each time point,
ii) Univariate analysis of variance,
iii) Univariate and multivariate analyses of time variables, and
iv) Mixed model methodology.

Separate analyses at each time point do not require special methods for repeated measures and do not directly address the objectives of examining and comparing trends over time. The other three approaches require special methodology and software. Development of statistical methods for repeated measures data has been an active area of research in the past two decades because of advancements in computing hardware and software. Enhancements in the SAS System reflect the advancements in methodology and hardware. In SAS System the GLM procedure enabled users to perform univariate analysis of variance but did not provide valid standard errors for most estimates. Moreover, conclusions derived from univariate analysis of variance are often invalid because the methodology does not adequately address the covariance structure of repeated measures. The REPEATED statement is now available to the SAS in the GLM procedure and Mixed procedure. PROC GLM provides both univariate and multivariate tests for repeated measures for one response. Another approach to analysis of repeated measures is via general mixed models. This approach can handle balanced as well as unbalanced or missing within-subject data, and it offers more options for modeling the within-subject covariance. The main drawback of the mixed models approach is that it generally requires iteration and, thus, may be less computationally efficient. The results provided by the REPEATED statement are based on univariate and multivariate analyses of contrast variables computed from the repeated measures variables. This approach basically bypassed the problems of covariance structure rather than addressing them directly. The REPEATED statement enabled users to obtain statistical tests for effects involving time trends. However, the tests were inefficient and the problem of incorrect standard errors remained. In addition, missing data on even one measure of an animal caused all the data for that animal to be ignored. Mixed procedure provided capabilities of mixed model methodology for analysis of repeated measures data. Use of mixed model methodology enabled the user to directly address the covariance structure and greatly enhanced the user's ability to analyze repeated measures data by providing valid standard errors and efficient statistical tests.

Here we shall illustrate the univariate and multivariate methods of analysis and their respective advantages and shortcomings. The statistical analysis methods illustrated focus on group (sex) comparisons at specific times, group comparisons averaged over times, and on changes over time in specific groups. Differences between groups (male and female) are computed at individual times and averaged across times.

Separate analyses at each time and the GLM REPEATED statement require the data to be organized in "multivariate mode." That is, there is one row per experimental unit in the data set, and the measurements at each time are considered separate response variables. The univariate ANOVA and MIXED procedure require that the data be organized in "univariate mode," that is, one row per experimental unit at each time.

We use the data obtained on body weight (kg) of pigs for the male and female. The body weights of pigs are collected at interval of 4 weeks since birth to 20 weeks of age and are given in Table - 1. Here the sex has two levels.

Table 1: Body weights of pigs maintained at Jabalpur

Animal No.	**Sex**	**Week**					
		0	**4**	**8**	**12**	**16**	**20**
1	**Male**	1	4.8	12.6	16	21	22.6
2	**Male**	1	4.2	7	10	14	22
3	**Male**	0.8	4	6	6.4	10	15
4	**Male**	0.8	4	6	9	13	21
5	**Male**	0.8	5	9.4	11	14	23
6	**Male**	0.8	3.2	7	10	15	22

7	**Male**	0.8	3.2	5.5	7.4	12	17
8	**Male**	0.8	3.4	7	8.7	12.4	19.2
9	**Female**	1	5.4	10	13	17.4	26.4
10	**Female**	1.2	4.8	12.6	16	20	21
11	**Female**	1	4.6	13	18	22	24
12	**Female**	0.8	4.2	8	11	13	18
13	**Female**	0.8	3.8	7	7.2	12	19
14	**Female**	1	5.4	11	14	19	22
15	**Female**	1	6	5.4	10	17	26.8
16	**Female**	1	3.4	7.8	10	13	17.8

Now the analysis of this data by using different methods with the use of software is given below:

I) Analysis at Individual Time Points

Analysis of data at each time point examines group effects separately at individual observation times and makes no statistical comparisons among times. This can be anlysed by using even in Microsoft Excel (easily available software). In it we make a file in Microsoft Excel by taking columns as the levels of the groups and then using Anova single factor command in Data Analysis command in Tools. This process is repeated for each time point.

No inference is drawn about trends over time, so this method is not truly a repeated measures analysis. Use of analysis at each time point is usually at a preliminary stage of data analysis and is not a preferred method because it does not address time effects. The only advantage in this method is that if we do not have any statistical software the data can be analyzed in Microsoft Excel.

II) Univariate ANOVA when the data follow a trend

Some of the repeated measures data such as growth, lactation data follow a trend. The analysis of such data can be done by fitting the appropriate such as linear, quadratic curves etc. on each of the animal. A set of estimates of parameters of these repeated data are estimated. These estimates are further analyzed to determine the effect of factors. The drawback of this method is that we are using the estimates of parameter which are not the true values and that may not be normally distributed.

III) Univariate Analysis of Variance Using the General Linear Model

Univariate analysis of variance (ANOVA), is the method most commonly applied to repeated measures data that makes comparisons between times. It treats the data as if they were from a split-plot design with the animals as whole-plot units and animals at particular times as sub-plot units. This approach also is referred to as a *split plot in time* analysis. If measurements have equal variance at all times, and if pairs of measurements on the same animal are equally correlated, regardless of the time lag between the measurements, then the univariate ANOVA is valid from a statistical point of view, and, in fact, yields an optimal method of analysis. However, measurements close in time are often more highly correlated than measures far apart in time, which will invalidate tests for effects involving time. For this procedure data is to be set in univariate mode

IV) Analysis of Contrast Variables

Contrast variables in repeated measures data are linear combinations of the responses over time for individual animals. A familiar example from basic statistical methodology is given by the orthogonal polynomials (Snedecor and Cochran, 1980), which represent linear, quadratic, cubic, etc., trends over time. Another example is the set of differences between responses at consecutive time points, that is, changes from time 1 to time 2, time 2 to time 3, and so forth. A set of contrast variables can be used to analyze trends over time and to make comparisons between times in repeated measures data. The original repeated measures data for each animal are transformed into a new set of variables given by a set of contrast variables. Then, multivariate and univariate analyses can be applied to these new variables. This provides a method for analyzing repeated measures data that evades some of the covariance structure problems that invalidate univariate ANOVA analyses, as discussed in the previous section.

V) Mixed Model Analysis

As noted above, analysis of repeated measures data requires special attention to the covariance structure due to the sequential nature of the data on each animal. Procedures discussed previously either avoid the issue (analysis of contrast variables) or ignore it (univariate analysis of variance). Ignoring the covariance issues may result in incorrect conclusions from the statistical analysis. Avoiding the issues may result in inefficient analyses, which is tantamount to wasting data. The general linear mixed model allows the capability to address the issue directly by modeling the covariance structure.

There are two basic steps in performing a repeated measures analysis using mixed model methodology. The first step is to model the covariance structure. The second step is to analyze time trends for groups by estimating and comparing means.

Measures on different animals are independent, so covariance concern is only with measures on the same animal. The covariance structure refers to variances at individual times and to correlation between measures at different times on the same animal. There are basically two aspects of the correlation. First, two measures on the same animal are correlated simply because they share common contributions from the animal. This is due to variation between animals. Second, measures on the same animal close in time are often more highly correlated than measures far apart in time. This is covariation within animals. Three different structures will be shown here and one will be chosen as best among the three. First, a structure known as compound symmetry (**CS**) will be fitted. This structure specifies that measures at all times have the same variance, and that all pairs of measures on the same animal have the same correlation. The implication is that the only aspect of the covariance between repeated measures is due to the animal contribution, irrespective of proximity of time.

Implications

Computer software is currently available that enables researchers to analyze repeated measures data using mixed model methodology. This methodology provides more valid and efficient statistical analyses of repeated measures. Implementation of this methodology requires the data analyst to model the variance and correlation structure of the data as a first step. Then, comparisons of groups and trends over time can be analyzed.

Illustration: An experiment was conducted to study the fruit (mango) weight for two types of pollination for four verities of mango with three replications. The experiment was planned with following parameters.

Factors	*Levels*	*Values*
Type of pollination	2	Selfed (1); Open pollination (2)

variety	4	Amarpali (1), Pusa (2), Arunima (3), Malika (4)
Replication	3	1, 2, 3

The data of fruit weight is given in Table-1.

Table-1: Fruit weight (g) at different time points

Pollination method	Variety	Replication	Time points (Weeks)			
			7	14	21	28
1	1	1	0.0325	0.2304	0.3580	0.412
1	1	2	0.0402	0.2364	0.449	0.521
1	1	3	0.046	0.2339	0.357	0.457
1	2	1	0.0243	0.224	0.426	0.512
1	2	2	0.0497	0.124	0.387	0.587
1	2	3	0.0406	0.1989	0.42	0.518
1	3	1	0.0348	0.1286	0.258	0.453
1	3	2	0.0335	0.0742	0.187	0.387
1	3	3	0.033	0.045	0.086	0.231
1	4	1	0.086	0.231	0.451	1.96
1	4	2	0.0533	0.249	0.449	1.345
1	4	3	0.0721	0.413	0.521	1.756
2	1	1	0.107	0.368	0.857	2.436
2	1	2	0.1225	0.326	0.511	1.957
2	1	3	0.089	0.14	0.355	2.594
2	2	1	0.0421	0.061	0.588	1.812
2	2	2	0.0515	0.078	0.677	1.571
2	2	3	0.0381	0.073	0.621	1.426
2	3	1	0.0413	0.0426	0.643	2.26
2	3	2	0.0312	0.0427	0.752	2.13
2	3	3	0.0317	0.047	0.632	2.563
2	4	1	0.1455	0.297	0.623	1.288
2	4	2	0.983	0.334	0.421	1.314
2	4	3	0.2286	0.308	0.545	1.074

Analyze the data for main effects of the factors and their interaction with time points using the repeated methodology.

ANALYSIS OF GROUPS OF EXPERIMENTS

1. Introduction

In large-scale experimental programmes it is necessary to repeat the trial of a set of treatments like varieties or manures at a number of places or in a number of seasons. The places where the trial is repeated are usually experimental stations located in the tract. The aim of repetition is to study the susceptibility of treatment effects to place variation. More generally, the aim of repetition is to find out treatments suitable for particular tracts in which case the trials are carried out simultaneous on a representative selection of sites.

Further, the purpose of the research carried out at experimental stations is to formulate the recommendations for the practitioners which consist of a population quite extensive either in space or time or both. Therefore, it becomes necessary to ensure that the results obtained from researches are valid for at least several places in the future and over a reasonably heterogeneous space.

A single experiment will precisely furnish information about only one place where the experiment is conducted and about the season in which the experiment is conducted. It has, thus, become a common practice to repeat an experiment at different places or over a number of occasions to obtain valid recommendations taking into account place to place variation or variation over time or both. In such cases of repeated experiments appropriate statistical procedures for a combined analysis of data would have to be followed by the analysis of individual experiments varying with their objectives. In combined analysis of data, the main points of interest would be

i) to estimate the average response to given treatments and

ii) to test consistency of the responses from place to place or occasion to occasion *i. e.* interaction of the treatment effects with places or years.

The utility and the significance of the estimates of average response depend on whether the response is consistence from place to place or changes with it, in other words on the absence or the presence of interaction.

The results of a set of trials may, therefore, be considered as belonging to one of the following four types:

i) the experimental errors are homogeneous and the interaction is absent,

ii) the experimental errors are homogeneous and the interaction is present,

iii) the experimental errors are heterogeneous and the interaction is absent, and

iv) the experimental errors are heterogeneous and the interaction is present.

The meaningfulness of average estimates of treatment responses would therefore, depend largely upon the absence of presence of this interaction analysis.

2. Analysis Procedure

For combined analysis or analysis for groups of Experiments following steps are to be followed

Step I: Construct an out line of combined analysis of variance over years or for places or environment, based on the basic design used. For example, the data of grain yield for four places, four treatments each treatment replicated five times is given in Table-1.

Step II: Perform usual Analysis of variance for the given data. Here the experiment conducted is in randomized complete block design. So perform analysis of four places separately for the four places. This may be done either in SAS, SPSS or EXCEL software.

Step III: We have *p* error mean squares that belongs to *p* RBD conducted and we have to test the homogeneity of variances. Now we have following two situations:

Situation I: When p = 2

In this situation, we apply *F*-test for testing the homogeneity of variances. Here null and alternate hypothesis are H_0: $\sigma_1^2 = \sigma_2^2$ and H_1: $\sigma_1^2 \neq \sigma_2^2$. Let Se_1^2 and Se_2^2 are the mean square errors (mse) for the two places. Then the value of *F* statistics will be $\mathrm{Se}_1^2/\mathrm{Se}_2^2$ and this value will be tested against the Table *F* value at $n_1$and n_2 degrees of freedom at 5 % level of significance, where $n_1$and n_2 are degrees of freedom (df) for error for the two places, respectively. If the calculated value of *F* is greater than tabulated *F* value then the null hypothesis of homogeneity of variance is rejected and the data is heterogeneous in different places, otherwise it is homogeneous.

Situation II: When p > 2

In this situation, we apply Bartlett's Chi-square test. Here null and alternate hypothesis are $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$ against the alternative hypothesis

H_1 : at least two of the σ_i^2's are not equal, where σ_i^2 is the error variance for i th place/ location.

Let Se_1^2, Se_2^2, ..., Se_p^2 are the mse of *p* locations respectively and n_1, n_2, ..., n_p are the df for p locations. The test statistics

$$\chi^2_{p-1} = \frac{\sum n_i \log \bar{s}_e^2 - \sum n_i \log s_{e_i}^2}{1 + \frac{1}{3(p-1)}(\sum \frac{1}{n_i} - \frac{1}{\sum n_i})}, \qquad \text{where} \quad \bar{s}_e^2 = \frac{\sum n_i s_{e_i}^2}{\sum n_i}$$

and if $n_i = n$

$$\chi^2_{p-1} = \frac{n[p \log \bar{s}_e^2 - \sum \log s_{e_i}^2]}{1 + \frac{(p+1)}{3np}}.$$

where χ^2_{p-1} follows χ^2 distribution with p - 1 degree of freedom.

If the calculated value of χ^2_{p-1} is greater than tabulated χ^2_{p-1} value at p-1df then the null hypothesis of homogeneity of variance is rejected and the data is heterogeneous in different places, otherwise it is homogeneous.

Step IV: If error variances are not homogeneous, then for performing the combined analysis of weighted least square is required, the weight being the reciprocals of the root mean square error. The weighted analysis is carried out by defining a new variable as newres = res/ root mean square. This transformation is similar to Aitken’s transformation. This new variable is thus homogeneous and thus combined analysis of variance can performed on this new variable. If error variance variances are homogeneous then there is no need to transform the data.

Step V: Now one can view the groups of experiments as a nested design with several factors nested within one another. The places/ locations are treated as big blocks, with the experiments nested within these. The combined analysis of data, therefore, can be done as that a nested design. For doing the analysis, the replication wise data of treatments at each place/ location provide useful information. An advantage of this analysis is that there is a further reduction in error sum of squares because one more source of variability is taken out from the experimental error thus reducing the experimental error. This may also lead to the reduction in the value of CV.

Step VI: Next step in the analysis is to test for the significance of place × treatment interaction. It can be seen that the question whether the interaction place × treatment is significant, that is whether the difference between treatments tend to vary from place to place can be settled by comparing the mean square for place × treatment with the estimate of error variance by the *F*-test. If the mean square is found to be non-significant it means interaction is absent. If this interaction is assumed to be non-existence, sum of squares for treatments × places and the error sum of squares can be pooled and a more precise estimate of error can be obtained for testing the significance of treatment differences. If, however, interaction is significant *i. e.* treatment effects are varying with places, then the appropriate mean square for testing the significance of treatments is the mean square due to place × treatment.

Exercise

Table-1: Data for grain yield (kg/ plot) with four treatments in five replications

Place	**Replication**					
	Treatment	**I**	**II**	**III**	**IV**	**V**
1	1	33.6	33.7	30.9	33.3	15.0
	2	34.0	27.2	46.2	36.7	11.6
	3	30.5	33.2	15.1	33.3	29.7
	4	30.8	14.4	14.2	9.5	12.0
2	1	28.8	28.8	35.2	41.6	43.2
	2	46.4	43.2	38.4	54.4	57.6
	3	35.2	32.0	32.0	25.6	33.6
	4	51.2	40.0	49.6	51.2	49.6
3	1	30.1	38.1	21.4	17.6	14.3
	2	36.1	18.3	38.0	31.0	26.6
	3	27.2	40.7	15.5	18.1	12.3
	4	37.8	54.5	13.2	18.1	7.3
4	1	23.8	48.8	19.5	28.8	34.4
	2	15.2	39.0	39.8	52.0	31.2
	3	40.2	52.0	33.0	41.2	35.0
	4	43.2	46.8	34.5	44.5	38.0

INCOMPLETE BLOCK DESIGNS

1. Introduction

Incomplete block designs are desirable when number of treatments to be tested is large and / or complete blocks are unavailable or inappropriate. These designs were introduced by Yates in order to eliminate heterogeneity to a greater extent as compared to a complete block design, when the number of treatments is large. The precision of the estimate of a treatment effect depends on the number of replications of the treatment - the larger the number of replications, the more is the precision. Similar is the case for the precision of estimate of the difference between two treatment effects. If a pair of treatment occurs together more number of times in the design, the difference between these two treatment effects can be estimated with more precision. To ensure equal or nearly equal precision of comparisons of different pairs of treatment effects, the treatments are so allocated to the experimental units in different blocks of equal sizes such that each treatment occurs at most once in a block and it has an equal number of replications and each pair of treatments has the same or nearly the same number of replications. When the number of replications of all pairs of treatments in a design is the same, then we have an important class of designs called **Balanced Incomplete Block** (BIB) designs and when there are unequal number of replications for different pairs of treatments, then the designs are called as **Partially Balanced Incomplete Block** (PBIB) designs. Another important class of incomplete block designs is lattice designs. Some of these are Balanced Incomplete Block (BIB) designs while others are Partially Balanced Incomplete Block (PBIB) designs.

2. Balanced Incomplete Block (BIB) Designs

A BIB design is an arrangement of v treatments in b blocks each of size k (<v) such that

(i) Each treatment occurs at most once in a block

(ii) Each treatment occurs in exactly r blocks

(iii) Each pair of treatments occurs together in exactly λ blocks.

Example 2.1: A BIB design for v = b = 5, r = k = 4 and λ = 3 in the following:

Blocks	
1	(1,2,3,4)
2	(1,2,3,5)
3	(1,2,4,5)
4	(1,3,4,5)
5	(2,3,4,5)

The symbols v, b, r, k, λ are called the parameters of the design. These parameters satisfy the relations

$$vr = bk \qquad \ldots(2.1)$$

and

$$\lambda(v-1) = r(k-1) \qquad \ldots(2.2)$$

A BIB design cannot exist unless (2.1) and (2.2) are satisfied. For instance, no design exists for v = b = 6 and r = k = 3 since, from (2.2) λ=6/5 is not an integer. However, these conditions are not sufficient for the existence of a BIB design. Even if both (2.1) and (2.2) are satisfied, it does not follow that such a design exists. For example, no BIB design exits for v = 15, b = 21, r = 7, k = 5, and λ = 2 even though both conditions are satisfied. In search of a criterion for the availability of a BIB design, Fisher proved that no design with b<v is possible.

Construction of BIB Designs

There is no single method of constructing all BIB designs. Solutions of many designs are still **unknown**. We describe below a few well known series of BIB designs.

2.1 Unreduced BIB Designs

These designs are obtained by taking all combinations of the v treatments k at a time. Therefore, the parameters of all unreduced BIB designs are:

$$v,\ k,\ b = {}^{v}C_{k},\ r = {}^{v-1}C_{k-1},\ \lambda = {}^{v-2}C_{k-2}$$

The BIB design for v = 5 treatments given in the previous section is an example of an unreduced BIB design in blocks of size 4.

Example 2.1: Let $v = 5$, $k = 3$, then $b = {}^{5}C_{3} = 10$, $r = {}^{4}C_{2} = 6$ and $\lambda = {}^{3}C_{1}$. The 10 blocks are:

Blocks	
1	(1,2,3)
2	(1,2,4)
3	(1,2,5)
4	(1,3,4)
5	(1,3,5)
6	(1,4,5)
7	(2,3,4)
8	(2,3,5)
9	(2,4,5)
10	(3,4,5)

These unreduced designs usually require a large number of blocks and replications so that the resulting designs will often be too large for practical purposes.

2.2 BIB Designs using MOLS

Before we describe the method, we explain the concept of mutually orthogonal Latin squares (MOLS) which will be used in the construction of BIB designs.

A Latin square of order s is an arrangement of s symbols in an s × s array such that each symbol occurs once in each row and once in each column of the array. For example, the following are 4 × 4 Latin squares of order 4 in symbols A, B, C, and D:

$$\begin{array}{cccc} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{array} \qquad \begin{array}{cccc} A & B & C & D \\ C & D & A & B \\ D & C & B & A \\ B & A & D & C \end{array} \qquad \begin{array}{cccc} A & B & C & D \\ D & C & B & A \\ B & A & D & C \\ C & D & A & B \end{array}$$

Two Latin squares are pairwise orthogonal if, when one square is superimposed on the other, each symbol of one Latin square occurs once with each symbol of the other square. Three or more squares are mutually orthogonal if they are pair-wise orthogonal. The three 4 × 4 Latin squares above are mutually orthogonal.

A complete set of s-1 mutually orthogonal Latin squares is known to exist for any $s = p^n$, where p is a prime number. Tables can be found in Fisher and Yates (1963). Now we describe the methods of constructing BIB designs using MOLS.

Suppose $v = s^2$ treatments are set out in an s × s array. A group of s blocks each of size s is obtained by letting the rows of the array represent blocks. Another group of s blocks is given by taking the columns of the array as blocks. Now suppose one of the orthogonal Latin squares is superimposed on to the array of treatments. A further group of s blocks is obtained if all treatments common to a particular symbol in the square are placed in a block. Each of the s-1 orthogonal squares produces a set of s blocks in this manner. The resulting design is a BIB design with parameters $v = s^2$, $b = s^2 + s$, $k = s$, $r = s + 1$, $\lambda = 1$.

Example 2.2: For $v = 3^2 = 9$ treatments a 3 × 3 array and a complete set of mutually orthogonal Latin squares of order 3 × 3 are :

$$\begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{array} \qquad \begin{array}{ccc} A & B & C \\ C & A & B \\ B & C & A \end{array} \qquad \begin{array}{ccc} A & B & C \\ B & C & A \\ C & A & B \end{array}$$

Four groups of 3 blocks are obtained from the rows, columns and the symbols of the two squares, as follows:

	Blocks		
	(1, 2, 3)		(1, 5, 9)
Rows	(4, 5, 6)	**First square**	(2, 6, 7)
	(7, 8, 9)		(3, 4, 8)
	(1, 4, 7)		(1, 6, 8)
Columns	(2, 5, 8)	**Second square**	(2, 4, 9)
	(3, 6, 9)		(3, 5, 7)

It can be checked that this is a BIB design with parameters $v = 9$, $b = 12$, $r = 4$, $k = 3$, and $\lambda = 1$.

2.3 Randomization Procedure

(i) Allot the treatment symbols (1,2,...,v) to the v treatments at random.

(ii) Allot the groups of k treatments to the b blocks at random.

(iii) Randomize the positions of the treatment numbers within each block.

2.4 Statistical Analysis

Consider the following model:

Observation = General mean + treatment effect + block effect + random error.

Random errors are assumed to be independently and identically distributed normally with mean zero and constant variance σ^2. On minimising the error sum of squares with respect to the parameters, we get a set of normal equations which can be solved to get the estimates of different contrasts of various treatment and block effects.

Now we compute

G = Grand total of observations
$\bar{y}$ = grand mean = G/n, where n= vr = bk = total number of observations
T_i = Sum of obervations for treatment i, (i=1,2,..., v)
B_j = Sum of observations in block j, (j=1,2,..., b)
CF= G^2/ n,
Q_i = adjusted ith treatment total
= T_i - (Sum of block totals in which treatment i occurs) / Block size (k)

A solution for the ith treatment effect is,

$\hat{\tau}$ = (k Q_i) / (λ v) (i = 1,2, ..., v)

Adjusted treatment mean for treatment i= ith treatment effects ($\hat{\tau}_i$) + grand mean ($\bar{y}$).

Various sums of squares can be obtained as follows:

(i) Total Sum of Squares (TSS) = Σ (observations)^2 - CF
(ii) Treatment Sum of Squares unadjusted (SST_u) = [Σ T_i^2] /r - CF
(iii) Block Sum of Squares unadjusted (SSB_U) = [Σ B_j^2] / k - CF
(iv) Treatments Sum of Squares adjusted (SST_A) = Σ $\hat{\tau}_i$ Q_i
(v) Error SS (SSE) = TSS - SSB_U - SST_A
(vi) Blocks sum of squares adjusted (SSB_A) = SST_A + SSB_U - SST_U

The analysis of variance for a BIB design is given below:

Table 2.1: ANOVA for a BIB (v, b, r, k, λ) Design

Source	DF	SS	MS	F
Treatment (unadj.)	v-1	SST_u		
Blocks (unadjusted)	b-1	SSB_u		
Treatments (adjusted)	v-1	SST_A	MST	MST/MSE
Blocks (adjusted)	b-1	SSB_A	MSB	MSB/MSE
Error	n-b-v+1	SSE	MSE	
Total	n-1	TSS		

Note: MST = SST_A / (v-1), MSB = SSB_A / (b-1) and MSE = SSE / (n -b- v + 1)
Coefficient of Variation = ($\sqrt{MSE}$ / $\bar{y}$) × 100

Standard error of difference between two adjusted treatment means = $[2k\,MSE/(\lambda v)]^{1/2}$.

C.D. = $t_{0.05} \times [2k\,MSE/(\lambda v)]^{1/2}$

3. Partially Balanced Incomplete Block (PBIB) Designs

BIB designs may not fit well to many experimental situations as these designs may not be available for all numbers of treatments and block sizes or may require a large number of replications. To overcome these difficulties PBIB designs were introduced. In these designs the variance of every estimated elementary contrast among treatment effects is not the same. The definition of PBIB designs is based on the association scheme.

Association Scheme

Given v treatment symbols 1,2,..,v, a relation satisfying the following conditions is called an m-class association scheme (m ≥2):

(i) Any two symbols are either 1^{st}, 2^{nd},..., or m^{th} associates; the relation of association being symmetric, *i.e*., if the symbol α is the i^{th} associate of β, then β is the i^{th} associate of α.

(ii) Each symbol α has n_i i^{th} associates, the number n_i being independent of α,

(iii) If any two symbols α and β are i^{th} associates, then the number of symbols that are j^{th} associates of α and k^{th} associate of β is p^i_{jk} and is independent of the pair of i^{th} associates α and β.

The numbers v, n_i and p^i_{jk} (i,j,k = 1,2,...,m) are called the parameters of the association scheme and satisfy the following relations:

$$\sum_{i=1}^{m} n_i = v-1$$

$$\sum_{k=1}^{m} p^i_{jk} = n_j - 1, \quad \text{if } i = j$$

$$= n_j , \quad \text{if } i \neq j$$

$$n_i p^i_{jk} = n_j p^j_{ik}$$

Example 3.1: Consider v=12 treatments denoted by numbers 1 to 12. Form 3 groups of 4 symbols each as follows: (1,2,3,4), (5,6,7,8), (9,10,11,12). We now define any two treatments as first associates if they belong to the same group, and second associates if they belong to the different groups. Here, n_1 = 3, n_2 = 8.

Definition: Given an association scheme with m classes (m ≥2) we have a PBIB design with m associate classes based on the association scheme, if the v treatment symbols can be arranged into b blocks, such that

(i) Every symbol occurs at most once in a block.

(ii) Every symbol occurs in exactly r blocks.

(iii) If two symbols are i^{th} associates, then they occur together in λ_i blocks, the number λ_i being independent of the particular pair of i^{th} associates α and β.

The numbers v, b, r, k, λ_i (i =1,2,...,m) are called the parameters of the design. It can be easily seen that

$$vr = bk \text{ and } \sum_{i=1}^{m} n_i\lambda_i = r(k-1).$$

It may be mentioned that as in the case of BIB designs, the complementary design of a PBIB with parameters v,b,r,k,λ_i is also a PBIB design having the same association scheme with the parameters v^*=v, b^*=b, r^*=b-r, k^*=v-k, λ_i^*=b-2r+λ_i.

PBIB designs can be broadly classified into (i) two-associate class PBIB designs (ii) three-associate class PBIB designs and (iii) higher associate class PBIB designs. Two-class association schemes and the two-associate PBIB designs have been extensively studied in the literature and are simple to use. As an illustration, we describe Group Divisible (GD) association scheme and the designs based on it.

3.1 GD Association Scheme

Let $v = mn$ symbols be arranged into m groups of n symbols each. A pair of symbols belonging to the same group is first associates $[n_1 = n\text{-}1]$ and a pair of symbols belonging to different groups is second associates $[n_2 = n(m\text{-}1)]$. A PBIB (2) design based on a GD scheme is called a GD design.

Method of Construction of Some GD Designs

Let D be a BIB design with parameters $v = m, b, r, k, \lambda$. Obtain a design D^* from D by replacing the i^{th} treatment (i=1,2,...,v) in D by n new treatment symbols $i_1, i_2, ..., i_n$. D^* is a group divisible design with the following parameters $v^* = mn$, $b^* = b$, $r^* = r$, $k^* = nk$, m, n, $\lambda_1 = r$, $\lambda_2 = \lambda$.

Example 3.1: Consider the following BIB design with parameters (4, 4, 3, 3, 2):

(1, 2, 3)
(1, 2, 4)
(1, 3, 4)
(2, 3, 4)

Replacing 1 by a, b; 2 by c, d; 3 by e, f and 4 by g, h, the following GD design with parameters $v = 8$, $b = 4$, $r = 3$, $k = 6$, $\lambda_1 = 3$, $\lambda_2 = 2$. is obtained:

(a, b, c, d, e, f)
(a, b, c, d, g, h)
(a, b, e, f, g, h)
(c, d, e, f, g, h)

3.2 Triangular association scheme and Design

3.2.1 Association scheme: Let there be n(n-1)/2 treatments arranged in a square array of size n such that the positions of the principal diagonal of the array are left blank, the n(n-1)/2 positions above the principal diagonal are filled up by the v treatment symbols and the positions below the principal diagonal are filled up by the v symbols in such a manner that the resultant arrangement is symmetrical about the principal diagonal.

Two treatments are first associates if they belong to same row or same column of the array and second associates, otherwise. Triangular scheme exists when $n \geq 5$ and here $v = n(n\text{-}1)/2$, $n \geq 5$, $n_1 = 2(n\text{-}2)$, $n_2 = (n\text{-}2)(n\text{-}3)/2$

$$P_1 = \begin{bmatrix} (n\text{-}2) & (n\text{-}3) \\ (n\text{-}3) & [(n\text{-}3)(n\text{-}4)]/2 \end{bmatrix} \quad P_2 = \begin{bmatrix} 4 & 2(n\text{-}4) \\ 2(n\text{-}4) & [(n\text{-}4)(n\text{-}5)]/2 \end{bmatrix}$$

Example 3.2.1: For n=5

*	1	2	3	4
1	*	5	6	7
2	5	*	8	9
3	6	8	*	10
4	7	9	10	*

Table 2.2 shows the various associates of all the treatments.

Table 2.2

Treatment	**1^{st} Associates**	**2^{nd} Associates**
1	2, 3,4,5, 6,7	8,9,10,
2	1,3,4,5, 8, 9	6,7, 10
3	1, 2,4, 6,8,10	5,7, 9
4	1,2,3,7,9,10,	5,6,8
5	1,2, 6,7, 8, 9	3,4,10
6	1,3,5,7, 8,10	2,4,9
7	1,4,5, 6, 9,10,	2,3,8
8	2,3,5, 6, 9,10	1,4,7
9	2,4,5,7, 8,10	1,3,6
10	3,4,6,7,8,9	1,2,5

3.2.2 Method of construction of Triangular designs: A two class association scheme is called triangular design if it is based on triangular association scheme**.** In a triangular association scheme, if we take each row as a block then the resultant design is triangular design with parameters v = n (n-1)/2, b=n, r=2, k=n-1, λ_1=1, λ_2= 0.

Example 3.2.2: Suppose n=5, giving rise to v=10 treatments as follows:

*	1	2	3	4
1	*	5	6	7
2	5	*	8	9
3	6	8	*	10
4	7	9	10	*

Taking each row as block, the following triangular design is obtained:

1	2	3	4
1	5	6	7
2	5	8	9
3	6	8	10
4	7	9	10

Here, v = 10, b=5, r=2, k=4, λ_1=1, λ_2= 0.

FACTORIAL EXPERIMENTS

1. Introduction

Factorial Experiments are experiments that investigate the effects of two or more factors or input parameters on the output response of a process. Factorial experiment design, or simply factorial design, is a systematic method for formulating the steps needed to successfully implement a factorial experiment. Estimating the effects of various factors on the output of a process with a minimal number of observations is crucial to being able to optimize the output of the process.

In a factorial experiment, the effects of varying the levels of the various factors affecting the process output are investigated. Each complete trial or replication of the experiment takes into account all the possible combinations of the varying levels of these factors. Effective factorial design ensures that the least number of experiment runs are conducted to generate the maximum amount of information about how input variables affect the output of a process.

For example, an experiment on rooting of cuttings involving two factors, each at two levels, such as two hormones at two doses, is referred to as a 2 x 2 or a 2^2 factorial experiment. Its treatments consist of the following four possible combinations of the two levels in each of the two factors.

Treatment number	**Treatment Combination**	
	Hormone	**Dose (ppm)**
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

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 example, 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. For a factorial experiment involving 5 clones, 4 espacements, and 3 weed-control methods, the total number of treatments would be 5 x 4 x 3 = 60. Thus, indiscriminate use of factorial experiments has to be avoided because of their large size, complexity, and cost. Furthermore, it is not wise to commit oneself to a large experiment at the beginning of the investigation when several small preliminary experiments may offer promising results. For example, a tree breeder has collected 30 new clones from a neighbouring country and wants to assess their reaction to the local environment. Because the environment is expected to vary in terms of soil fertility, moisture levels, and so on, the ideal experiment would be one that tests the 30 clones in a factorial experiment involving such other variable factors as fertilizer, moisture level, and population density. Such an experiment, however, becomes extremely large as factors other than clones are added. Even if only one factor, say nitrogen or fertilizer with three levels were included, the number of treatments would increase from 30 to 90. Such a large experiment would mean difficulties in financing, in obtaining an adequate experimental area, in controlling soil heterogeneity, and so on. Thus, the more practical approach would be to test the 30 clones first in a single-factor experiment, and then use the results to select a few clones for further studies in more detail. For example, the initial single-factor experiment may show that only five clones are outstanding enough to warrant further testing. These five clones could then be put into a factorial experiment with three levels of nitrogen, resulting in an experiment with 15 treatments rather than the 90 treatments needed with a factorial experiment with 30 clones.

The amount of change produced in the process output for a change in the 'level' of a given factor is referred to as the 'main effect' of that factor. Table 1 shows an example of a simple factorial experiment involving two factors with two levels each. The two levels of each factor may be denoted as 'low' and 'high', which are usually symbolized by '-' and '+' in factorial designs, respectively.

Table 1. A Simple 2-Factorial Experiment

	A (-)	**A (+)**
B (-)	20	40
B (+)	30	52

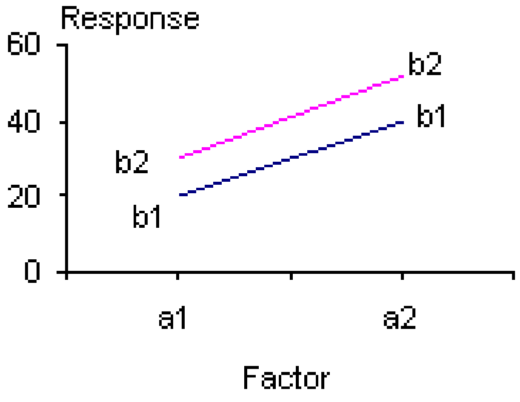
The main effect of a factor is basically the 'average' change in the output response as that factor goes from '-' to '+'. Mathematically, this is the average of two numbers: 1) the change in output when the factor goes from low to high level as the other factor stays low, and 2) the change in output when the factor goes from low to high level as the other factor stays high.

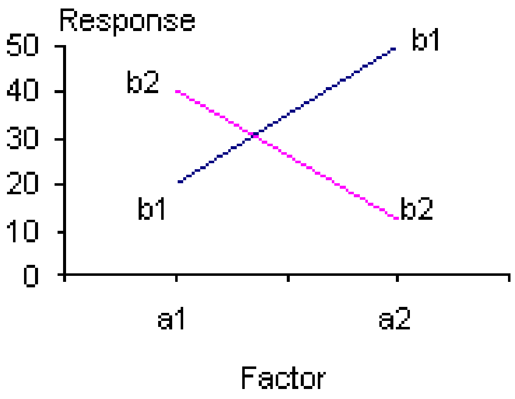
In the example in Table 1, the output of the process is just 20 (lowest output) when both A and B are at their '-' level, while the output is maximum at 52 when both A and B are at their '+' level. The main effect of A is the average of the change in output response when B stays '-' as A goes from '-' to '+', or (40-20) = 20, and the change in output response when B stays '+' as A goes from '-' to '+', or (52-30) = 22. The main effect of A, therefore, is equal to 21.

Similarly, the main effect of B is the average change in output as it goes from '-' to '+' , i.e., the average of 10 and 12, or 11. Thus, the main effect of B in this process is 11. Here, one can see that the factor A exerts a greater influence on the output of process, having a main effect of 21 versus factor B's main effect of only 11. It must be noted that aside from 'main effects', factors can likewise result in 'interaction effects.' Interaction effects are changes in the process output caused by two or more factors that are interacting with each other. Large interactive effects can make the main effects insignificant, such that it becomes more important to pay attention to the interaction of the involved factors than to investigate them individually. In Table 1, as effects of A (B) is not same at all the levels of B (A) hence, A and B are interacting.

Thus, **interaction** is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance through out all the levels of other factors OR the factors are said to interact if the effect of one factor changes as the levels of other factor(s) changes.

Graphical representation of lack of interaction between factors and interaction between factors are shown below. In case of two parallel lines, the factors are non-interacting.





If interactions exist which is fairly common, we should plan our experiments in such a way that they can be estimated and tested. It is clear that we cannot do this if we vary only one factor at a time. For this purpose, we must use multilevel, multifactor experiments.

The running of factorial combinations and the mathematical interpretation of the output responses of the process to such combinations is the essence of factorial experiments. It allows to understand which factors affect the process most so that improvements (or corrective actions) may be geared towards these.

We may define factorial experiments as experiments in which the effects (main effects and interactions) of more then one factor are studied together. In general if there are ‘n’ factors, say, F_1, F_2,..., F_n and i^{th} factor has s_i levels, i=1,...,n, then total number of treatment combinations is $\prod_{i=1}^{n} s_i$. Factorial experiments are of two types.

Experiments in which the number of levels of all the factors are same i.e all s_i’s are equal are called **symmetrical factorial experiments** and the experiments in at least two of the s_i‘s are different are called as **asymmetrical factorial experiments**. Factorial experiments provide an opportunity to study not only the individual effects of each factor but also there interactions. They have the further advantage of economising on experimental resources. When the experiments are conducted factor by factor much more resources are required for the same precision than when they are tried in factorial experiments.

2. Experiments with Factors Each at Two Levels

The simplest of the symmetrical factorial experiments are the experiments with each of the factors at 2 levels. If there are ‘n’ factors each at 2 levels, it is called as a 2^n factorial where the power stands for the number of factors and the base the level of each factor. Simplest of the symmetrical factorial experiments is the 2^2 factorial experiment i.e. 2 factors say A and B each at two levels say 0 (low) and 1 (high). There will be 4 treatment combinations which can be written as

00	$= a_0 b_0$	= 1;	A and B both at first (low) levels
10	$= a_1 b_0$	= a ;	A at second (high) level and B at first (low) level

01 $= a_0 b_1 =$ b ; A at first level (low) and B at second (high) level
11 $= a_1 b_1 =$ ab; A and B both at second (high) level.

In a 2^2 factorial experiment wherein r replicates were run for each combination treatment, the main and interactive effects of A and B on the output may be mathematically expressed as follows:

A = [ab + a - b - (1)] / 2r; (main effect of factor A)
B = [ab + b - a - (1)] / 2r; (main effect of factor B)
AB = [ab + (1) - a - b] / 2r; (interactive effect of factors A and B)

where r is the number of replicates per treatment combination; a is the total of the outputs of each of the r replicates of the treatment combination a (A is 'high and B is 'low); b is the total output for the n replicates of the treatment combination b (B is 'high' and A is 'low); ab is the total output for the r replicates of the treatment combination ab (both A and B are 'high'); and (1) is the total output for the r replicates of the treatment combination (1) (both A and B are 'low’).

Had the two factors been independent, then [ab + (1) - a - b] / 2n will be of the order of zero. If not then this will give an estimate of interdependence of the two factors and it is called the interaction between A and B. It is easy to verify that the interaction of the factor B with factor A is BA which will be same as the interaction AB and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect of factor B, a contrast of the treatment totals is orthogonal to each of A and AB.

Table 2. Two-level 2-Factor Full-Factorial

RUN	Comb.	M	A	B	AB
1	(1)	+	-	-	+
2	a	+	+	-	-
3	b	+	-	+	-
$4 = 2^2$	ab	+	+	+	+

Consider the case of 3 factors A, B, C each at two levels (0 and 1) i.e. 2^3 factorial experiment. There will be 8 treatment combinations which are written as

000 $= a_0 b_0 c_0$ = (1); A, B and C all three at first level
100 $= a_1 b_0 c_0$ = a ; A at second level and B and C at first level
010 $= a_0 b_1 c_0$ = b ; A and C both at first level and B at second level
110 $= a_1 b_1 c_0$ = ab; A and B both at second level and C is at first level.
001 $= a_0 b_0 c_1$ = c ; A and B both at first level and C at second level.
101 $= a_1 b_0 c_1$ = ac; A and C at second level, B at first level
011 $= a_0 b_1 c_1$ = bc; A is at first level and B and C both at second level
111 $= a_1 b_1 c_1$ = abc; A, B and C all the three at second level

In a three factor experiment there are three main effects A, B, C; 3 first order or two factor interactions AB, AC, BC; and one second order or three factor interaction ABC.

Table 3. Two-level 3-Factor Full-Factorial Experiment Pattern

RUN	Comb.	M	A	B	AB	C	AC	BC	ABC
1	(1)	+	-	-	+	-	+	+	-
2	a	+	+	-	-	-	-	+	+
3	b	+	-	+	-	-	+	-	+
4	ab	+	+	+	+	-	-	-	-
5	c	+	-	-	+	+	-	-	+
6	ac	+	+	-	-	+	+	-	-
7	bc	+	-	+	-	+	-	+	-
$8 = 2^3$	abc	+	+	+	+	+	+	+	+

$$\text{Main effect A} = \frac{1}{4}\{[abc] -[bc] +[ac] -[c] + [ab] -[b] + [a] -[1]\}$$

$$= \frac{1}{4}(a-1)\,(b+1)\,(c+1)$$

$$AB = \frac{1}{4}\ [(abc)-(bc) -(ac) +c) - (ab) - (b) - (a)+ (1)\ ]$$

$$ABC = \frac{1}{4}\ [\ (abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)\ ]$$

or equivalently,

$$AB = \frac{1}{4}\ (a-1)\,(b-1)\,(c+1)$$

$$ABC = \frac{1}{4}\ (a-1)\,(b-1)\,(c-1)$$

The method of representing the main effect or interaction as above is due to Yates and is very useful and quite straightforward. For example, if the design is 2^4 then

$$A = (1/2^3)\ [\ (a-1)\,(b+1)\,(c+1)\,(d+1)\ ]$$
$$ABC = (1/2^3)\ [\ (a-1)\,(b-1)\,(c-1)\,(d+1)]$$

In case of a 2^n factorial experiment, there will be 2^n (=v) treatment combinations with 'n' main effects, $\binom{n}{2}$ first order or two factor interactions, $\binom{n}{3}$ second order or three factor interactions, $\binom{n}{4}$ third order or four factor interactions and so on , $\binom{n}{r}$, (r-1)th order or r factor interactions and $\binom{n}{n}$ (n-1)th order or n factor interaction. Using these v treatment combinations, the experiment may be laid out using any of the suitable experimental designs viz. completely randomised design or block designs or row-column designs, etc.

Steps for Analysis

1. The Sum of Squares (S.S.) due to treatments, replications [in case randomised block design is used], due to rows and columns (in case a row-column design has been used), total S.S. and error S.S. is obtained as per established procedures. No replication S.S. is required in case of a completely randomised design.
2. The treatment sum of squares is divided into different components viz. main effects and interactions each with single d.f. The S.S. due to these factorial effects is obtained by dividing the squares of the factorial effect total by $r.2^n$. For obtaining 2^n-1 factorial effects in a 2^n factorial experiment, the 'n' main effects is obtained by giving the positive signs to those treatment totals where the particular factor is at second level and minus to others and dividing the value so obtained by $r.2^{n-1}$, where r is the number of replications of the treatment combinations. All interactions can be obtained by multiplying the corresponding coefficients of main effects.

   For a 2^2 factorial experiment, the S.S. due to a main effect or the interaction effect is obtained by dividing the square of the effect total by 4r. Thus,

   S.S. due to main effect of A = $[A]^2/ 4r$, with 1 d.f.

   S.S. due to main effect of B = $[B]^2/ 4r$, with 1 d.f

   S.S. due to interaction AB = $[AB]^2/ 4r$, with 1 d.f.

3. Mean squares (M.S) is obtained by dividing each S.S. by corresponding degrees of freedom.
4. After obtaining the different S.S.'s, the usual Analysis of variance (ANOVA) table is prepared and the different effects are tested against error mean square and conclusions drawn.
5. Standard errors (S.E.'s) for main effects and two factor interactions:

   S.E of difference between main effect means = $\sqrt{\frac{2MSE}{r.2^{n-1}}}$

   S.E of difference between A means at same level of B=S.E of difference between B means at same level of A= $\sqrt{\frac{2MSE}{r.2^{n-2}}}$

In general,

S.E. for difference between means in case of a r-factor interaction = $\sqrt{\frac{2MSE}{r.2^{n-r}}}$

The critical differences are obtained by multiplying the S.E. by the student's t value at α% level of significance at error degrees of freedom.

The ANOVA for a 2^2 factorial experiment with r replications conducted using a RCBD is as follows:

ANOVA

Sources of Variation	**DF**	**S.S.**	**M.S.**	**F**
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^2-1=3	SST	MST=SST/3	MST/MSE
A	1	SSA=[A]^2/4r	MSA=SSA	MSA/MSE
B	1	SSB=[B]^2/4r	MSB=SSB	MSB/MSE
AB	1	SSAB=[AB]^2/4r	MSAB=SSAB	MSAB/MSE
Error	(r-1)(2^2-1) =3(r-1)	SSE	MSE=SSE/3(r-1)	
Total	r.2^2-1=4r-1	TSS		

ANOVA for a 2^3-factorial experiment conducted in RCBD with r replications is given by

ANOVA

Sources of Variation	**DF**	**SS**	**MS**	**F**
Between Replications	r-1	SSR	MSR=SSR/(r-1)	MSR/MSE
Between treatments	2^3 -1=7	SST	MST=SST/7	MST/MSE
A	1	SSA	MSA=SSA	MSA/MSE
B	1	SSB	MSB=SSB	MSB/MSE
C	1	SSC	MSC=SSC	MSC/MSE
AB	1	SSAB	MSAB=SSAB	MSAB/MSE
AC	1	SSAC	MSAC=SSAC	MSAC/MSE
BC	1	SSBC	MSBC=SSBC	MSBC/MSE
ABC	1	SSABC	MSABC=SSABC	MSABC/MSE
Error	(r-1)(2^3-1) =7(r-1)	SSE	MSE=SSE/7(r-1)	
Total	r.2^3-1=8r-1	TSS		

Similarly ANOVA table for a 2^n factorial experiment can be made.

3. Experiments with Factors Each at Three Levels

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels, the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment increases too rapidly with them. Consider two factors A and B, each at three levels say 0, 1 and 2 (3^2-factorial experiment). The treatment combinations are

00	= a_0b_0	= 1	; A and B both at first levels
10	= a_1b_0	= a	; A is at second level and B is at first level
20	= a_2b_0	= a^2	; A is at third level and b is at first level
01	= a_0b_1	= b	; A is at first level and B is at second level
11	= a_1b_1	= ab	; A and B both at second level
21	= a_2b_1	= a^2b	; A is at third level and B is at second level
02	= a_0b_2	= b^2	; A is at first level and B is at third level
12	= a_1b_2	= ab^2	; A is at second level and B is at third level
22	= a_2b_2	= a^2b^2	; A and B both at third level

Any standard design can be adopted for the experiment.

The main effects A, B can respectively be divided into linear and quadratic components each with 1 d.f. as A_L, A_Q, B_L and B_Q. Accordingly AB can be partitioned into four components as A_L B_L , A_L B_Q, A_Q B_L, A_Q B_Q.

The coefficients of the treatment combinations to obtain the above effects are given as

Treatment Totals→ Factorial Effects ↓	**[1]**	**[a]**	**[a^2]**	**[b]**	**[ab]**	**[a^2b]**	**[b^2]**	**[ab^2]**	**[a^2b^2]**	**Divisor**
M	+1	+1	+1	+1	+1	+1	+1	+1	+1	9r=r×3^2
A_L	-1	0	+1	-1	0	+1	-1	0	+1	6r=r×2× 3
A_Q	+1	-2	+1	+1	-2	+1	+1	-2	+1	18r=6×3
B_L	-1	-1	-1	0	0	0	+1	+1	+1	6r=r×2×3
A_L B_L	+1	0	-1	0	0	0	-1	0	+1	4r=r×2×2
A_Q BL	-1	+2	-1	0	0	0	+1	-2	+1	12r=r×6×2
B_Q	+1	+1	+1	-2	-2	-2	+1	+1	+1	18r=r×3×6
A_L B_Q	-1	0	+1	+2	0	-2	-1	0	+1	12r=r×2×6
A_Q B_Q	+1	-2	+1	-2	+4	-2	+1	-2	+1	36r=r×6×6

The rule to write down the coefficients of the linear (quadratic) main effects is to give a coefficient as +1 (+1) to those treatment combinations containing the third level of the corresponding factor, coefficient as 0(-2) to the treatment combinations containing the second level of the corresponding factor and coefficient as -1(+1) to those treatment combinations containing the first level of the corresponding factor. The coefficients of the treatment combinations for two factor interactions are obtained by multiplying the corresponding coefficients of two main effects. The various factorial effect totals are given as

$$[A_L] = +1[a^2b^2]+0[ab^2] -1[b^2]+1[a^2b]+0[ab] -1[b]+1[a^2]+0[a] -1[1]$$

$$[A_Q] = +1[a^2b^2] -2[ab^2]+1[b^2]+1[a^2b] -2[ab]+1[b]+1[a^2] -2[a]+1[1]$$

$$[B_L] = +1[a^2b^2]+1[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2] -1[a] -1[1]$$

$$[A_LB_L] = +1[a^2b^2]+0[ab^2] -1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+0[a] -1[1]$$

$$[A_QB_L] = +1[a^2b^2] -2[ab^2]+1[b^2]+0[a^2b]+0[ab]+0[b] -1[a^2]+2[a] -1[1]$$

$$[B_Q] = +1[a^2b^2]+1[ab^2]+1[b^2] -2[a^2b] -2[ab] -2[b] -1[a^2] -1[a] -1[1]$$

$$[A_LB_Q]= +1[a^2b^2]+0[ab^2] -1[b^2] -2[a^2b]+0[ab]+2[b]+1[a^2]+0[a] -1[1]$$

$$[A_QB_Q] = +1[a^2b^2] -2[ab^2]+1[b^2] -2[a^2b]+4[ab] -2[b]+1[a^2] -2[a]+1[1]$$

Factorial effects are given by
$A_L = [A_L]/r.3$ $A_Q= [A_Q]/r.3$ $B_L = [B_L]/r.3$ $A_LB_L = [A_LB_L]/r.3$

$A_QB_L = [A_QB_L]/r.3$ $B_Q = [B_Q]/r.3$ $A_LB_Q = [A_LB_Q]/r.3$ $A_QB_Q = [A_QB_Q]/r.3$

The sum of squares due to various factorial effects is given by

$$SSA_L = \frac{[A_L]^2}{r.2.3}; \quad SSA_q = \frac{[A_Q]^2}{r.6.3}; \quad SSB_L = \frac{[B_L]^2}{r.3.2}; \quad SSA_LB_L = \frac{[A_LB_L]^2}{r.2.2};$$

$$SSA_QB_L = \frac{[A_QB_L]^2}{r.6.2}; \quad SSB_Q= \frac{[B_Q]^2}{r.3.6}; \quad SSA_LB_Q = \frac{[A_LB_Q]^2}{r..2.6}; \quad SSA_QB_Q = \frac{[A_QB_Q]^2}{r.6.6};$$

If a RBD is used with r-replications then the outline of analysis of variance is

ANOVA

Sources of Variation	**D.f**	**SS**	**MS**
Between Replications	r-1	SSR	MSR=SSR/(r-1)
Between treatments	3^2-1=8	SST	MST=SST/8
A	2	SSA	MSA=SSA/2
A_L	1	SSA_L	$MSA_L= SSA_L$
A_Q	1	SSA_Q	$MSA_Q=SSA_Q$
B	2	SSB	MSB=SSB/2
B_L	1	SSB_L	$MSB_L= SSB_L$
B_Q	1	SSB_Q	$MSB_Q=SSB_Q$
AB	4	SSAB	MSAB=SSAB/2
A_LB_L	1	SSA_LB_L	$MSA_LB_L=SSA_LB_L$
A_QB_L	1	SSA_QB_L	$MSA_QB_L=SSA_QB_L$
A_LB_Q	1	SSA_LB_Q	$MSA_LB_Q=SSA_LB_Q$
A_QB_Q	1	SSA_QB_Q	$MSA_QB_Q=SSA_QB_Q$
Error	(r-1)(3^2-`1) =8(r-1)	SSE	MSE=SSE/8(r-1)
Total	r.3^2-1=9r-1	TSS	

In general, for n factors each at 3 levels, the sum of squares due to any linear (quadratic) main effect is obtained by dividing the square of the linear (quadratic) main effect total by $r.2.3^{n-1}$($r.6.3^{n-1}$). Sum of squares due to a 'p' factor interaction is given by taking the square of the total of the particular interaction component divided by r.(a_1 a_2 ...a_p). 3^{n-p}, where a_1, a_2,...,a_p are taken as 2 or 6 depending upon the linear or quadratic effect of particular factor.

4. Confounding in Factorial Experiments

When the number of factors and/or levels of the factors increase, the number of treatment combinations increase very rapidly and it is not possible to accommodate all these treatment combinations in a single homogeneous block. For example, a 2^5 factorial would have 32 treatment combinations and blocks of 32 plots are quite big to ensure homogeneity within them. A new technique is therefore necessary for designing experiments with a large number of treatments. One such device is to take blocks of size less than the number of treatments and have more than one block per replication. The treatment combinations are then divided into as many groups as the number of blocks per replication. The different groups of treatments are allocated to the blocks.

There are many ways of grouping the treatments into as many groups as the number of blocks per replication. It is known that for obtaining the interaction contrast in a factorial experiment where each factor is at two levels, the treatment combinations are divided into two groups. Such two groups representing a suitable interaction can be taken to form the contrasts of two blocks each containing half the total number of treatments. In such case the contrast of the interaction and the contrast between the two block totals are given by the same function. They are, therefore, mixed up and can not be separated. In other words, the interaction has been confounded with the blocks. Evidently the interaction confounded has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by taking one or more interaction contrasts identical with block contrasts is known as **confounding**. Preferably only higher order interactions, that is, interactions with three or more factors are confounded, because their loss is immaterial. As an experimenter is generally interested in main effects and two factor interactions, these should not be confounded as far as possible.

When there are two or more replications, if the same set of interactions are confounded in all the replications, confounding is called **complete** and if different sets of interaction are confounded in different replications, confounding is called **partial**. In complete confounding all the information on confounded interactions are lost. But in partial confounding, the confounded interactions can be recovered from those replications in which they are not confounded.

Advantages of Confounding

It reduces the experimental error considerably by stratifying the experimental material into homogeneous subsets or subgroups. The removal of the variation among incomplete blocks (freed from treatments) within replicates results in smaller error mean square as compared with a RBD, thus making the comparisons among some treatment effects more precise.

Disadvantages of Confounding

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partial or complete) on certain relatively unimportant interactions.
- The confounded contrasts are replicated fewer times than are the other contrasts and as such there is loss of information on them and they can be estimated with a lower degree of precision as the number of replications for them is reduced.
- An indiscriminate use of confounding may result is complete or partial loss of information on the contrasts or comparisons of greatest importance. As such the experimenter should confound only those treatment combinations or contrasts which are of relatively less or of importance at all.
- The algebraic calculations are usually more difficult and the statistical analysis is complex, especially when some of the units (observations) are missing.

Confounding in 2^3 Experiment

Although 2^3 is a factorial with small number of treatment combinations but for illustration purpose, this example has been considered. Let the three factors be A, B, C each at two levels.

Factorial Effects → **Treat. Combinations ↓**	A	B	C	AB	AC	BC	ABC
(1)	-	-	-	+	+	+	-
(a)	+	-	-	-	-	+	+
(b)	-	+	-	-	+	-	-
(ab)	+	+	-	+	-	-	-
(c)	-	-	+	+	-	-	+
(ac)	+	-	+	-	+	-	-
(bc)	-	+	+	-	-	+	-
(abc)	+	+	+	+	+	+	+

The various factorial effects are as follows:

A = (abc) + (ac) + (ab) + (a) - (bc) - (c) - (b) - (1)
B = (abc) + (bc) + (ab) + (b) - (ac) - (c) - (a) - (1)
C = (abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)
AB = (abc) + (c) + (ab) + (1) - (bc) - (ac) - (b) - (a)
AC = (abc) + (ac) + (b) + (1) - (bc) - (c) - (ab) - (a)
BC = (abc) + (bc) + (a) + (1) - (ac) - (c) - (ab) - (b)
ABC = (abc) + (c) + (b) + (a) - (bc) - (ac) - (ab) - (1)

Let the highest order interaction ABC be confounded and we decide to use two blocks of 4 units (plots) each per replicate.

Thus in order to confound the interaction ABC with blocks all the treatment combinations with positive sign are allocated at random in one block and those with negative signs in the other block. Thus the following arrangement gives ABC confounded with blocks and hence we loose information on ABC.

	Replication I			
Block 1:	(1)	(ab)	(ac)	(bc)
Block 2:	(a)	(b)	(c)	(abc)

It can be observed that the contrast estimating ABC is identical to the contrast estimating block effects.

The other six factorial effects viz. A, B, C, AB, AC, BC each contain two treatments in block 1 (or 2) with the positive signs and two with negative sign so that they are orthogonal with block totals and hence these differences are not influenced among blocks and can thus be estimated and tested as usual without any difficulty. Whereas for confounded interaction, all the treatments in one group are with positive sign and in the other with negative signs.

Similarly if AB is to be confounded, then the two blocks will consists of

Block 1	(abc)	(c)	(ab)	(1)
Block 2	(bc)	(ac)	(b)	(a)

Here AB is confounded with block effects and cannot be estimated independently whereas all other effects A, B, C, AC, Bc and ABC can be estimated independently.

When an interaction is confounded in one replicate and not in another, the experiment is said to be partially confounded. Consider again 2^3 experiment with each replicate divided into two blocks of 4 units each. It is not necessary to confound the same interaction in all the replicates and several factorial effects may be confounded in one single experiment. For example, the following plan confounds the interaction ABC, AB, BC and AC in replications I, II, III and IV respectively.

Rep. I		**Rep. II**		**Rep. III**		**Rep. IV**	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Block 7**	**Block 8**
(abc)	(ab)	(abc)	(ac)	(abc)	(ab)	(abc)	(ab)
(a)	(ac)	(c)	(bc)	(bc)	(ac)	(ac)	(bc)
(b)	(bc)	(ab)	(a)	(a)	(b)	(b)	(a)
(c)	(1)	(1)	(b)	(1)	(c)	(1)	(c)

In the above arrangement, the main effects A, B and C are orthogonal with block totals and are entirely free from block effects. The interaction ABC is completely confounded with blocks in replicate 1, but in the other three replications the ABC is orthogonal with blocks and consequently an estimate of ABC may be obtained from replicates II, III and IV. Similarly it is possible to recover information on the other confounded interactions AB (from I, III, IV), BC (from I, II, IV) and AC (from I, II, III). Since the partially confounded interactions are estimated from only a portion of the observations, they are determined with a lower degree of precision than the other effects.

For carrying out the statistical analysis, the various factorial effects and their S.S. are estimated in the usual manner with the modification that for **completely confounded** interactions neither the S.S due to confounded interaction is computed nor it is included in the ANOVA table. The confounded component is contained in the (2p-1) degrees of freedom (D.f.) (in case of p replicates) due to blocks. The partitioning of the d.f for a 2^3 completely confounded factorial is as follows.

Source of Variation	**D.f**
Blocks	2p-1
A	1
B	1
C	1
AB	1
AC	1
BC	1
Error	6(p-1)
Total	8p-1

In general for a 2^n completely confounded factorial in p replications, the different d.f's are given as follows:

Source of Variation	D.f
Replication	p-1
Blocks within replication	p(2^{n-r}-1)
Treatments	2^n-1-(2^{n-r}-1)
Error	By subtraction
Total	p2^n-1

The treatment d.f has been reduced by 2^{n-r}-1 as this is the total d.f confounded per block.

In case of partial confounding, we can estimate the effects confounded in one replication from the other replication in which it is not confounded. In (2^n, 2^r) factorial experiment with p replications, following is the splitting of d.f's.

Source of Variation	D.f
Replication	p-1
Blocks within replication	p(2^{n-r}-1)
Treatments	2^n-1
Error	By subtraction
Total	p2^n-1

The S.S. for confounded effects are to be obtained from those replications only in which the given effect is not confounded.

5. Fractional Factorial

In a factorial experiment, as the number of factors to be tested increases, the complete set of factorial treatments may become too large to be tested simultaneously in a single experiment. A logical alternative is an experimental design that allows testing of only a fraction of the total number of treatments. A design uniquely suited for experiments involving large number of factors is the fractional factorial. It provides a systematic way of selecting and testing only a fraction of the complete set of factorial treatment combinations. In exchange, however, there is loss of information on some pre-selected effects. Although this information loss may be serious in experiments with one or two factors, such a loss becomes more tolerable with large number of factors. The number of interaction effects increases rapidly with the number of factors involved, which allows flexibility in the choice of the particular effects to be sacrificed. In fact, in cases where some specific effects are known beforehand to be small or unimportant, use of the fractional factorial results in minimal loss of information.

In practice, the effects that are most commonly sacrificed by use of the fractional factorial are high order interactions - the four-factor or five-factor interactions and at times, even the three-factor interaction. In almost all cases, unless the researcher has prior information to indicate otherwise one should select a set of treatments to be tested so that all main effects and two-factor interactions can be estimated.

In forestry research, the fractional factorial is to be used in exploratory trials where the main objective is to examine the interactions between factors. For such trials, the most appropriate fractional factorials are those that sacrifice only those interactions that involve more than two factors.

With the fractional factorial, the number of effects that can be measured decreases rapidly with the reduction in the number of treatments to be tested. Thus, when the number of effects to be measured is large, the number of treatments to be tested, even with the use of fractional factorial, may still be too large. In such cases, further reduction in the size of the experiment can be achieved by reducing the number of replications. Although the use of fractional factorial without replication is uncommon in forestry experiments, when fractional factorial is applied to exploratory trials, the number of replications required can be reduced to the minimum.

Another desirable feature of fractional factorial is that it allows reduced block size by not requiring a block to contain all treatments to be tested. In this way, the homogeneity of experimental units within the same block can be improved. A reduction in block size is, however, accompanied by loss of information in addition to that already lost through the reduction in number of treatments.

6. Practicals on Factorial Experiments

Exercise 1: Analyse the data of a 2^3 factorial experiment conducted using a RCBD with three replications. The three factors were the fertilizers viz. Nitrogen (N), Phosphorus (P) and Potassium (K). The purpose of the experiment is to determine the effect of different kinds of fertilizers on crop yield. The yields under 8 treatment combinations for each of the three randomized blocks are given below:

Block- I

npk	(1)	k	np	p	n	nk	pk
450	101	265	373	312	106	291	391

Block- II

p	nk	k	np	(1)	npk	pk	n
324	306	272	338	106	449	407	89

Block- III

p	npk	nk	(1)	n	k	pk	np
323	471	334	87	128	279	423	324

Analysis

Step 1: The data is arranged in the following table:

Blocks ↓	**Treatment Combinations**								**Total**
	(1)	**n**	**p**	**np**	**k**	**nk**	**pk**	**npk**	
B_1	101	106	312	373	265	291	391	450	2289 (B_1)
B_2	106	89	324	338	272	306	407	449	2291 (B_2)
B_3	87	128	323	324	279	334	423	471	2369 (B_3)
Total	294	323	959	1035	816	931	1221	1370	6949

	(T_1)	(T_2)	(T_3)	(T_4)	(T_5)	(T_6)	(T_7)	(T_8)	(G)

Grand Total G = 6949,

Number of observations (n) =24 = $(r.2^n)$

Correction Factor (C.F.) = $\dfrac{G^2}{n} = \dfrac{(6949)^2}{24} = 2012025.042$

Total S.S. (TSS) = Sum $(\text{Obs.})^2$ - C.F = $(101^2 + 106^2 + ... + 449^2 + 471^2)$ - C.F = 352843.958

Block (Replication) S.S (SSR) = $\sum_{j=1}^{r} \dfrac{B_j^2}{2^3} - \text{C.F} = \dfrac{\left[(2289)^2 + (2291)^2 + (2369)^2\right]}{8} - \text{C.F}$

= 520.333

Treatment S.S.(SST) = $\sum_{i=1}^{v} \dfrac{T_i^2}{r} - \text{C.F}$

$$= \frac{(294)^2 + (323)^2 + (959)^2 + (1035)^2 + (816)^2 + (931)^2 + (1221)^2 + (1370)^2}{3} - \text{C.F}$$

$$= \frac{7082029}{3} - 2012025.042 = 348651.2913$$

Error S.S.(SSE) =Total S.S - Block S.S - Treatment S.S

= 352843.958 - 520.333 - 348651.2913 = 3672.3337

Step 2: Main effects totals and interactions totals are obtained as follows:

N = [npk]- [pk] +[nk] - [k] +[np] - [p]+[n]- [1] = 369

P = [npk]+ [pk] - [nk] - [k] +[np] +[p] -[n]- [1] = 2221

K = [npk]+ [pk] +[nk] +[k] - [np] - [p] -[n]- [1] = 1727

NP = [npk] - [pk] - [nk] +[k] +[np] - [p] -[n]+[1] = 81

NK = [npk] - [pk] +[nk] - [k] - [np]+ [p] -[n]+[1] = 159

PK = [npk]+ [pk] - [nk] - [k] - [np]- [p]+[n]+[1] = -533

NPK = [npk] - [pk] - [nk] +[k] - [np]+ [p]+[n]- [1] =-13

Factorial effects = $\dfrac{\text{Factorial effect Total}}{r.2^{n-1}(=12)}$

Factorial effect SS = $\dfrac{(\text{Factorial effect Total})^2}{r.2^n(=24)}$

Here Factorial Effects

N=30.75, P=185.083, K=143.917, NP=6.75, NK=13.25, PK=-44.417, NPK=-1.083

SS due to N = 5673.375

SS due to P = 205535.042

SS due to K =124272.0417

SS due to NP = 273.375

SS due to NK=1053.375

SS due to PK = 11837.0417

SS due to NPK=7.04166.

Step 3: M.S. is obtained by dividing S.S.'s by respective degrees of freedom.

ANOVA

Sources of Variation	**DF**	**SS**	**MS**	**F**
Replications	r-1=2	520.333	260.167	0.9918
Treatments	2^3-1=7	348651.291	49807.3273	189.8797*
N	(s-1)=1	5673.375	5673.375	21.6285*
P	1	205535.042	205535.042	783.5582*
K	1	124272.042	124272.042	473.7606*
NP	1	273.375	273.375	1.0422
NK	1	1053.375	1053.375	4.0158
PK	1	11837.041	11837.041	45.1262*
NPK	1	7.0412	7.0412	0.02684
Error	(r-1) (2^n-1)=14	3672.337	262.3098	
Total	r.2^n-1=23	352843.958		

*indicates significance at 5% level of significance

Step 5: S.E of difference between main effect means = $\sqrt{\frac{\text{MSE}}{\text{r}.2^{n-2}}}$ =8.098

S.E of difference between N means at same level of P or K = S.E of difference between P (or K) means at same level of N =S.E of difference between P means at same level of K = S. E. of difference between K means at same level of P = $\sqrt{\frac{\text{MSE}}{\text{r}.2^{n-3}}}$ = 11.4523. $t_{0.05}$ at 14 d.f. = 2.145. Accordingly critical differences (C.D.) can be calculated.

Exercise 2: The data on mean maximum culm height of *Bambusa arundinacea* tested with two levels of spacing (Factor A, 10 m x 10 m and 12 m x 12m) and three levels of age at planting (Factor B, 6, 12 and 24 months) laid out in RCBD with three replications is given below.

Treatment combination	**Maximum culm height of a clump (cm)**		
	Rep. I	**Rep. II**	**Rep. III**
a_1b_1	46.50	55.90	78.70
a_1b_2	49.50	59.50	78.70

a_1b_3	127.70	134.10	137.10
a_2b_1	49.30	53.20	65.30
a_2b_2	65.50	65.00	74.00
a_2b_3	67.90	112.70	129.00

ANOVA of a 2 x 3 Factorial Experiment in RCBD

Source of variation	Df	SS	MS	F
Replication	2	2040.37	1020.187	8.60*
Treatment	5	14251.87	2850.373	24.07*
A	1	12846.26	6423.132	3.45
B	2	408.98	408.980	54.12*
AB	2	996.62	498.312	4.20*
Error	10	1186.86	118.686	
Total	17	17479.10		

*Significant at 5% level.

The result indicates that the main effect of factor A (spacing) is not significant at the 5% level of significance. The analysis shows a significant interaction between spacing and age, indicating that the effect of age vary with the change in spacing.

Exercise 3: A 3^2 experiment was conducted to study the effects of the two factors Nitrogen (N) and Phosphorus (P) (each at three levels 0, 1, 2) on sugar beets. Two replications of nine plots each were used. The table shows the plan and the percentage of sugar (approximated to nearest whole number).

Replication	Treatment		% of sugar
	N	P	
I	0	1	14
	2	0	15
	0	0	16
	2	1	15
	0	2	16
	1	2	18
	1	1	17
	1	0	19
	2	2	17
II	1	2	20
	1	0	19
	1	1	17
	0	0	18
	2	1	19

	0	1	16
	0	2	16
	2	2	19
	2	0	16

Analyse the data.

Analysis

Step 1: Sum of squares for replications, treatments and total sum of squares is obtained by arranging the data in a Replication x Treatment table as follows:

Rep.	**Treatment Combinations**									**Total**
	1 **00**	**n** **10**	**n^2** **20**	**p** **01**	**np** **11**	**n^2p** **21**	**p^2** **02**	**np^2** **12**	**n^2p^2** **22**	
1	16	19	15	14	17	15	16	18	17	147 (R_1)
2	18	19	16	16	17	19	16	20	19	160 (R_2)
Total	34 (T_1)	38 (T_2)	31 (T_3)	30 (T_4)	34 (T_5)	34 (T_6)	32 (T_7)	38 (T_8)	36 (T_9)	307 (G)

Grand Total = 307,

No. of observations (N) = $r.3^2$ =18

Correction Factor (C.F.) = $\frac{(307)^2}{18}$ =5236.0556

Total S.S.(TSS) = Sum(observation)^2-C.F. = $16^2+18^2+...+17^2+19^2$-5236.0556 = 48.9444

Replication SS (SSR) = $\frac{R_1^2+R_2^2}{9}-\text{C.F.}=\frac{147^2+160^2}{9}-5236.0556$ = 9.3888

Treatment SS (SST) = $\frac{\text{Sum(treatment totals)}^2}{r}-\text{C.F.}$

$$= \frac{34^2+38^2+...+38^2+36^2}{2}-5236.0556 = 32.4444$$

Error SS = Total SS - Replication SS - Treatment SS = 7.1112

Step 2: Obtain various factorial effects totals

$[N_L]$ =+1[n^2p^2]+0[np^2] -1[p^2]+1[n^2p]+0[np] -1[p]+1[n^2]+0[n] -1[1] = 5

$[N_Q]$ =+1[n^2p^2] -2[np^2]+1[p^2]+1[n^2p] -2[np]+1[p]+1[n^2] -2[n]+1[1] =-23

$[P_L]$ =+1[n^2p^2]+1[np^2]+1[p^2]+0[n^2p]+0[np]+0[p] -1[n^2] -1[n] -1[1] = 3

$[N_LP_L]$ =+1[n^2p^2]+0[np^2] -1[p^2]+0[n^2p]+0[np]+0[p] -1[n^2]+0[n] -1[1] = 7

$[N_QP_L]$ =+1[n^2p^2] -2[np^2]+1[p^2]+0[n^2p]+0[np]+0[p] -1[n^2]+2[n] -1[1] = 3

$[P_Q] \quad =+1[n^2p^2]+1[np^2]+1[p^2] -2[n^2p] -2[np] -2[p] -1[n^2] -1[n] -1[1] \qquad = 13$

$[N_LP_Q] =+1[n^2p^2]+0[np^2] -1[p^2] -2[n^2p]+0[np]+2[p]+1[n^2]+0[n] -1[1] \qquad =-7$

$[N_QP_Q]=+1[n^2p^2] -2[np^2]+1[p^2] -2[n^2p]+4[np] -2[p]+1[n^2] -2[n]+1[1] \qquad =-11$

Step 3: Obtain the sum of squares due to various factorial effects

$SSN_L = \frac{[N_L]^2}{r.2.3} = \frac{5^2}{12} = 2.0833;$ $\qquad SSN_Q = \frac{[N_Q]^2}{r.6.3} = \frac{(-23)^2}{36} = 14.6944;$

$SSP_L = \frac{[P_L]^2}{r.3.2} = \frac{3^2}{12} = 0.7500;$ $\qquad SSN_LP_L = \frac{[N_LP_L]^2}{r.2.2} = \frac{7^2}{8} = 6.1250;$

$SSN_QP_L = \frac{[N_QP_L]^2}{r.6.2} = \frac{3^2}{24} = 0.375;$ $\qquad SSP_Q= \frac{[P_Q]^2}{r.3.6} = \frac{13^2}{36} = 4.6944;$

$SSN_LP_Q = \frac{[N_LP_Q]^2}{r..2.6} = \frac{(-7)^2}{24} = 2.0417;$ $SSN_QP_Q = \frac{[N_QP_Q]^2}{r.6.6} = \frac{(-11)^2}{72} = 1.6806;$

ANOVA

Sources of Variation	DF	SS	MS	F
Between Replications	1	9.3888	9.3888	10.5623*
Between treatments	8	32.4444	4.0555	4.5624*
N	2	16.7774	8.3887	9.4371*
N_L	1	2.0833	2.0833	2.3437
N_Q	1	14.6944	14.6944	16.5310*
P	2	5.4444	2.7222	3.0624
P_L	1	0.7500	0.7500	0.8437
P_Q	1	4.6944	4.6944	5.2811
NP	4	10.2223	2.5556	2.875
N_LP_L	1	6.1250	6.1250	6.8905*
N_QP_L	1	0.3750	0.3750	0.4219
N_LP_Q	1	2.0417	2.0417	2.2968
N_QP_Q	1	1.6806	1.6806	1.8906
Error	8	7.1112	0.8889	
Total	17	48.9444		

*indicates the significance at 5%

SPLIT AND STRIP PLOT DESIGNS

1. Split Plot Design

1.1 Introduction

In conducting experiments, sometimes some factors have to be applied in larger experimental units while some other factors can be applied in comparatively smaller experimental units. Further some experimental materials may be rare while the other experimental materials may be available in large quantity or when the levels of one (or more) treatment factors are easy to change, while the alteration of levels of other treatment factors are costly, or time-consuming. One more point may be that although two or more different factors are to be tested in the experiment, one factor may require to be tested with higher precision than the others. In all such situations, a design called the split plot design is adopted.

A split plot design is a design with at least one blocking factor where the experimental units within each block are assigned to the treatment factor levels as usual, and in addition, the blocks are assigned at random to the levels of a further treatment factor. The designs have a nested blocking structure. In a block design, the experimental units are nested within the blocks, and a separate random assignment of units to treatments is made within each block. In a split plot design, the experimental units are called split-plots (or sub-plots), and are nested within whole plots (or main plots).

In split plot design, plot size and precision of measurement of effects are not the same for both factors, the assignment of a particular factor to either the main plot or the sub-plot is extremely important. To make such a choice, the following guidelines are suggested:

Degree of Precision- For a greater degree of precision for factor B than for factor A, assign factor B to the sub-plot and factor A to the main plot e.g. a plant breeder who plans to evaluate ten promising rice varieties with three levels of fertilization, would probably wish to have greater precision for varietal comparison than for fertilizer response. Thus, he would designate variety as the sub-plot factor and fertilizer as the main plot factor. Or, an agronomist would assign variety to main plot and fertilizer to sub-plot if he wants greater precision for fertilizer response than variety effect.

Relative Size of the Main effects- If the main effect of one factor (A) is expected to be much larger and easier to detect than that of the other factor (B), factor A can be assigned to the main plot and factor B to the sub-plot. This increases the chance of detecting the difference among levels of factor B which has a smaller effect.

Management Practices- The common type of situation when the split plot design is automatically suggestive is the difficulties in the execution of other designs, i.e. practical execution of plans. The cultural practices required by a factor may dictate the use of large plots. For practical expediency, such a factor may be assigned to the main plot e.g. in an experiment to evaluate water management and variety, it may be desirable to assign water mangement to the main plot to minimize water movement between adjacent plots, facilitate the simulation of the water level required, and reduce border effects. Or, if ploughing is one of the factors of interest, then one cannot have different depths of ploughing in different plots scattered randomly apart.

1.2 Randomization and Layout

There are two separate randomization processes in a split plot design – one for the main plot and another for the sub-plot. In each replication, main plot treatments are first randomly assigned to the main plots followed by a random assignment of the sub-plot treatments within each main plot. This procedure is followed for all replications. A possible layout of a split plot experiment with four main plot treatments(a=4), three sub-plot treatments(b=3), and four replications(r=4) is given below:

Rep. I				Rep. II				Rep. III				Rep. IV			
b_1	b_3	b_2	b_2	b_3	b_1	b_2	b_1	b_3	b_1	b_2	b_3	b_2	b_3	b_3	b_1
b_3	b_2	b_1	b_3	b_1	b_2	b_1	b_3	b_2	b_3	b_3	b_2	b_1	b_2	b_1	b_2
b_2	b_1	b_3	b_1	b_2	b_3	b_3	b_2	b_1	b_2	b_1	b_1	b_3	b_1	b_2	b_3
a_4	a_2	a_1	a_3	a_1	a_4	a_2	a_3	a_3	a_2	a_4	a_1	a_1	a_4	a_3	a_2

The above layout has the following important features – • The size of the main plot is *b* times the size of the sub-plot, • Each main plot treatment is tested *r* times whereas each sub-plot treatment is tested *ar* times, thus the number of times a sub-plot treatment is tested will always be larger than that for the main plot and is the primary reason for more precision for the sub-plot treatments relative to the main plot treatments.

This concept of splitting each plot may be extended further to accommodate the application of additional factors. An extension of this design is called the split-split plot design where the sub-plot is further divided to include a third factor in the experiment. The design allows for 3 different levels of precision associated with the 3 factors. That is, the degree of precision associated with the main factor is lowest, while the degree of precision associated with the sub-sub plot is the highest.

1.3 Model

The model for simple split plot design is

$$Y_{ijk} = \mu + \rho_i + \tau_j + \delta_{ij} + \beta_k + (\tau\beta)_{jk} + \varepsilon_{ijk}$$

for i = 1,2, …,r, j = 1,2, …,a, k = 1,2, …b,
where,

Y_{ijk} : observation corresponding to kth level of sub-plot factor(B), jth level of main plot factor(A) and the ith replication.

μ : general mean

ρ_i : ith block effect

τ_j : jth main plot treatment effect

β_k : kth sub-plot treatment effect

$(\tau\beta)_{jk}$: interaction between jth level of main-plot treatment and the kth level of sub-plot treatment

The error components δ_{ij} and ε_{ijk} are independently and normally distributed with means zero and respective variances σ^2_δ and σ^2_ε.

1.4 Analysis

Whole-Plot analysis:

This part of the analysis is based on comparisons of whole-plot totals:

- The levels of A are assigned to the whole plots within blocks according to a randomized complete block design, and so the sum of squares for A needs no block adjustment. There are *a* –1 degrees of freedom for A, so the sum of squares is given by

  $$\mathrm{ssA} = \sum_j y_{.j.}^2 / \mathrm{rb} - y_{...}^2 / \mathrm{rab}$$

  [The "dot" notation means "add over all values of the subscript replaced with a dot"]

- There are *r* –1 degrees of freedom for blocks, giving a block sum of squares of

  $$\mathrm{ssR} = \sum_i y_{i..}^2 / \mathrm{ab} - y_{...}^2 / \mathrm{rab}$$

- There are *a* whole plots nested within each of the *r* blocks, so there are, in total, *r*(*a* -1) whole-plot degrees of freedom. Of these, *a* –1 are used to measure the effects of A leaving (*r* –1)(*a* –1) degrees of freedom for whole-plot error. Equivalently, this can be obtained by the subtraction of the block and A degrees of freedom from the whole-plot total degrees of freedom i.e. (*ra* –1) –(*r* –1) –(*a* –1) = (*r* –1)(*a* –1).

  So, the whole plot error sum of squares, is obtained as

  $$\mathrm{ssE}_1 = \sum_i \sum_j y_{ij.}^2 / \mathrm{b} - y_{...}^2 / \mathrm{rab} - \mathrm{ssR} - \mathrm{ssA}$$

- The whole plot error mean square $\mathrm{msE}_1 = \mathrm{ssE}_1$ / (*r* –1)(*a* –1), is used as the error estimate to test the significance of whole plot factor(A).

Sub-plot analysis:

This part of the analysis is based on the observations arising from the split-plots within whole plots:

- There are *rab* –1 total degrees of freedom, and the total sum of squares is

  $$\mathrm{sstot} = \sum_i \sum_j \sum_k y_{ijk}^2 - y_{...}^2 / \mathrm{rab}$$

- Due to the fact that all levels of B are observed in every whole plot as in a randomized complete block design, the sum of squares for B needs no adjustment for whole plots, and is given by -

  $\mathrm{ssB} = \sum_k y_{..k}^2 / \mathrm{ra} - y_{...}^2 / \mathrm{rab}$, corresponding to *b* –1 degrees of freedom.

- The interaction between the factors A and B is also calculated as part of the split-plot analysis. Again, due to the complete block structure of both the whole-plot design and the split-plot design, the interaction sum of squares needs no adjustment for blocks. The number of interaction degrees of freedom is (*a* –1)(*b* –1), and the sum of squares is

  $$\mathrm{ss(AB)} = \sum_j \sum_k y_{.jk}^2 / \mathrm{r} - y_{...}^2 / \mathrm{rab} - \mathrm{ssA} - \mathrm{ssB}$$

- Since there are *b* split plots nested within the *ra* whole plots, there are, in total, *ra*(*b* –1) split-plot degrees of freedom. Of these, *b* –1 are used to measure the main effect of B, and (*a* –1)(*b* –1) are used to measure the AB interaction, leaving *ra*(*b*–1) – (*b*–1) – (*a*–1)(*b*–1) = *a*(*r*–1)(*b*–1) degrees of freedom for error. Equivalently, this can be obtained by subtraction of the whole plot, B, and AB degrees of freedom from the total i.e. (*rab* –1) – (*ra* – 1) – (*b* – 1) – (*a* –1)(*b* –1) = *a*(*r* –1)(*b* –1).

The split-plot error sum of squares can be calculated by subtraction:
$ssE_2 = sstot - ssR - ssA - ssE_1 - ssB - ss(AB)$.

- The split-plot error mean square $msE_2 = ssE_2 / a(r-1)(b-1)$ is used as the error estimate in testing the significance of split-plot factor(B) and interaction(AB).

- The analysis of variance table is outlined as follows:

ANOVA

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F**
Whole plot analysis				
Replication	r-1	ssR	-	-
Main plot treatment(A)	a-1	ssA	msA	msA/msE_1
Main plot error(E_1)	(r-1)(a-1)	ssE_1	$msE_1 = E_a$	
Sub-plot analysis				
Sub-plot treatment(B)	b-1	ssB	msB	msB/msE_2
Interaction (AxB)	(a-1)(b-1)	ss(AB)	ms(AB)	$ms(AB)/msE_2$
Sub-plot error(E_2)	a(r-1)(b-1)	ssE_2	$msE_2 = E_b$	
Total	rab-1	sstot		

1.5 Standard Errors and Critical Differences

Estimate of S.E. of difference between two main plot treatment means = $\sqrt{\frac{2E_a}{rb}}$

Estimate of S.E. of difference between two sub-plot treatment means = $\sqrt{\frac{2E_b}{ra}}$

Estimate of S.E. of difference between two sub-plot treatment means
at the same level of main plot treatment = $\sqrt{\frac{2E_b}{r}}$

Estimate of S.E. of difference between two main plot treatment
means at the same or different levels of sub-plot treatment = $\sqrt{\frac{2[(b-1)E_b + E_a]}{rb}}$

Critical difference is obtained by multiplying the $S.E_d$ by $t_{5\%}$ table value for respective error d.f. for (i), (ii) & (iii). For (iv), as the standard error of mean difference involves two error terms, we use the following equation to compute the weighted t values:

$$t = \frac{(b-1)E_b t_b + E_a t_a}{(b-1)E_b + E_a}$$

where t_a and t_b are t-values at error d.f. (E_a) and error d.f.(E_b) respectively.

Example 1: In a study carried by agronomists to determine if major differences in yield response to N fertilization exist among different varieties of jowar, the main plot treatments were three varieties of jowar (V_1: CO-18, V_2: CO-19 and V_3: C0-22), and the sub-plot treatments were N rates of 0, 30, and 60 Kg/ha. The study was replicated four times, and the data gathered for the experiment are shown in Table 1.

Table 1: Replication-wise yield data.

		N rate, Kg/ha		
Replication	Variety	0	30	60
			Yield, kg per plot	
I	V_1	15.5	17.5	20.8
	V_2	20.5	24.5	30.2
	V_3	15.6	18.2	18.5
II	V_1	18.9	20.2	24.5
	V_2	15.0	20.5	18.9
	V_3	16.0	15.8	18.3
III	V_1	12.9	14.5	13.5
	V_2	20.2	18.5	25.4
	V_3	15.9	20.5	22.5
IV	V_1	12.9	13.5	18.5
	V_2	13.5	17.5	14.9
	V_3	12.5	11.9	10.5

Analyze the data and draw conclusions.

Steps of analysis:

- Calculate the replication totals (R), and the grand total (G) by first constructing a table for the replication × variety totals shown in Table 1.1, and then a second table for the variety × nitrogen totals as shown in Table 1.2.

Table 1.1 Replication × Variety (RA) - table of yield totals.

	Variety			
Replication	V_1	V_2	V_3	Rep.Total(R)
I	53.8	75.2	52.3	181.3
II	63.6	54.4	50.1	168.1
III	40.9	64.1	58.9	163.9
IV	44.9	45.9	34.9	125.7
Variety Total(A)	203.2	239.6	196.2	
Grand Total(G)				639.0

Table 1.2 Variety × Nitrogen (AB) - table of yield totals.

	Variety			
Nitrogen	V_1	V_2	V_3	Nitrogen Total(B)
N_0	60.2	69.2	60.0	189.4
N_1	65.7	81.0	66.4	213.1
N_2	77.3	89.4	69.8	236.5

- Compute the various sums of squares for the main plot analysis by first computing the correction factor:

$$\text{C.F.} = \frac{G^2}{rab} = \frac{(639)^2}{4\times3\times3} = 11342.25$$

Total S.S. (sstot) = [$(15.5)^2 + (20.5)^2 + \ldots + (10.5)^2$] - C.F.
= 637.97

$$\text{Replication S.S. (ssR)} = \frac{\sum R^2}{ab} - \text{C.F.}$$

$$= \frac{(181.3)^2 + (168.1)^2 + (163.9)^2 + (125.7)^2}{3\times3} - 11342.25$$

= 190.08

$$\text{S.S. due to Variety (ssA)} = \frac{\sum A^2}{rb} - \text{C.F.}$$

$$= \frac{(203.2)^2 + (239.6)^2 + (196.2)^2}{4\times3} - 11342.25$$

= 90.487

$$\text{Main plot error S.S. } (\text{ssE}_1) = \frac{\sum (RA)^2}{b} - \text{C.F.} - \text{ssR} - \text{ssA}$$

$$= \frac{(53.8)^2 + (63.6)^2 + \ldots + (34.9)^2}{3} - 11342.25 - 190.08 - 90.487$$

= 174.103

- Compute the various sums of squares for sub-plot analysis:

$$\text{S.S. due to Nitrogen (ssB)} = \frac{\sum B^2}{ra} - \text{C.F.}$$

$$= \frac{(189.4)^2 + (213.1)^2 + (236.5)^2}{4\times3} - 11342.25$$

= 92.435

$$\text{S.S. due to Interaction } (A \times B) = \frac{\sum (AB)^2}{r} - \text{C.F.} - \text{ssA} - \text{ssB}$$

$$= \frac{(60.2)^2 + (65.7)^2 + \ldots + (69.8)^2}{4} - 11342.25 - 90.487 - 92.435$$

= 9.533

Sub-plot error S.S. (ssE_2) = Total S.S. – All other sum of squares
= 637.97 – (190.08 + 90.487 + 174.103 + 92.435 +9.533)
= 81.332

- Calculate the mean square for each source of variation by dividing the S.S. by its corresponding degrees of freedom and compute the F value for each effect that needs to be tested, by dividing each mean square by the corresponding error mean square, as shown in Table 1.3.

Table 1.3 ANOVA

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F**
Replication	3	190.08	63.360	
Variety(A)	2	90.487	45.243	1.56^{ns}
Error(a)	6	174.103	$29.017(E_a)$	
Nitrogen(B)	2	92.435	46.218	10.23^{**}
Variety×Nitrogen (A×B)	4	9.533	2.383	<1
Error(b)	18	81.332	$4.518\ (E_b)$	
Total	35	637.97		

^{ns} – not significant, ** - significant at 1% level.

- Compute the coefficient of variation for the main plot and sub-plot as:

$$\text{cv(a)} = \frac{\sqrt{E_a}}{\text{G.M.}} \times 100, \text{ and } \quad \text{cv(b)} = \frac{\sqrt{E_b}}{\text{G.M.}} \times 100 \text{ respectively.}$$

- Compute standard errors and to make specific comparisons among treatment means compute respective critical differences only when F-tests show significance differences and interpret.

- Conclusion: There was no significant difference among variety means. Yield was significantly affected by nitrogen. However, the interaction between N rate and variety was not significant. All the varieties gave significant response to 30 kg N/ha as well as to 60 kg N/ha.

2. Strip Plot Design

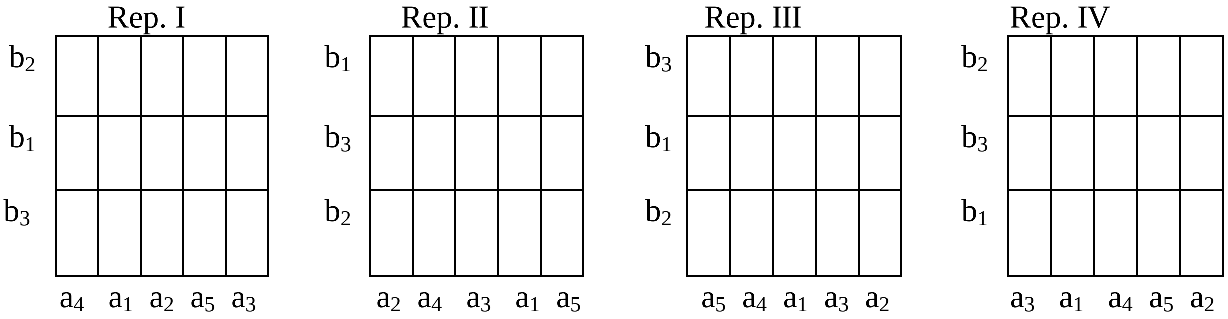
2.1 Introduction

Sometimes situation arises when two factors each requiring larger experimental units are to be tested in the same experiment, e.g., suppose four levels of spacing and three levels of methods of ploughing are to be tested in the same experiment. Here both the factors require large experimental units. If the combinations of the two factors at all possible levels are allotted in a R.B.D. in the normal way, the experimental plots shall have to be very large thereby bringing heterogeneity. So, it will not be appropriate. On the other hand if one factor (spacing) is taken in main plots and other factor (methods of ploughing) is taken in sub-plots within main plots, the sub-plots shall have to be large enough. Hence split plot design also will not be appropriate. In such situations a design called Strip plot design is adopted.

The strip plot is a 2-factor design that allows for greater precision in the measurement of the interaction effect while sacrificing the degree of precision on the main effects. The experimental area is divided into three plots, namely the vertical-strip plot, the horizontal-strip plot, and the intersection plot. We allocate factors A and B, respectively, to the vertical and horizontal-strip plots, and allow the intersection plot to accommodate the interaction between these two factors. As in the split plot design, the vertical and the horizontal plots are perpendicular to each other. However, in the strip plot design the relationship between the vertical and horizontal plot sizes is not as distinct as the main and sub-plots were in the split plot design. The sub-plot treatments instead of being randomized independently within each main plot as in the case of split plot design are arranged in strips across each replication. The intersection plot, which is one of the characteristics of the design, is the smallest in size.

2.2 Randomization and Layout

In this design each block is divided into number of vertical and horizontal strips depending on the levels of the respective factors. Let A represent the vertical factor with *a* levels, B represent the horizontal factor with *b* levels and *r* represent the number of replications. To layout the experiment, the experimental area is divided into *r* blocks. Each block is divided into *b* horizontal strips and *b* treatments are randomly assigned to these strips in each of the *r* blocks separately and independently. Then each block is divided into *a* vertical strips and *a* treatments are randomly assigned to these strips in each of the *r* blocks separately and independently. A possible layout of a strip plot experiment with *a* =5 (a_1, a_2, a_3, a_4, and a_5), *b* =3 (b_1, b_2, and b_3) and four replications is given below:



The strip plot design sacrifices precision on the main effects of both the factors in order to provide higher precision on the interaction which will generally be more accurately determined than in either randomised blocks or simple split plot design. Consequently this design is not recommended unless practical considerations necessitate its use or unless the interaction is the principle object of study.

2.3 Model

The model for strip plot design is

$$Y_{ijk} = \mu + \rho_i + \alpha_j + (\rho\alpha)_{ij} + \beta_k + (\rho\beta)_{ik} + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

for i = 1,2, …,r, j = 1,2, …,a, k = 1,2, …b,

where,

Y_{ijk} : observation corresponding to j^{th} level of factor A, k^{th} level of factor B and i^{th} replication

μ : general mean

ρ_i : i^{th} block effect

α_j : effect of j^{th} level of factor A

β_k : effect of k^{th} level of factor B

$(\alpha\beta)_{jk}$: interaction between j^{th} level of factor A and the k^{th} level of factor B

The error components $(\rho\alpha)_{ij}$, $(\rho\beta)_{ik}$ and ε_{ijk} are independently and normally distributed with means zero and respective variances σ^2_a, σ^2_b, and σ^2_ε.

2.4 Analysis

In statistical analysis separate estimates of error are obtained for main effects of the factor, A and B and for their interaction AB. Thus there will be three error mean squares applicable for testing the significance of main effects of the factors and their interaction separately.

Suppose 4 levels of spacings (A) and 3 levels of methods (B) of ploughing are to be tested in the same experiment. Each replication is divided into 4 strips vertically and into 3 strips horizontally. In the vertical strips the four different levels of spacings are allotted randomly and in the horizontal strips three methods of ploughing are allotted randomly. Let there be 4 replications(R). The analysis of variance is carried out in three parts viz. vertical strip analysis, horizontal strip analysis and interaction analysis as follows:

- Form spacing × replication (A × R) table of yield totals and from this table compute the S.S. due to replication, S.S. due to spacings and S.S. due to interaction - Replication × Spacing i.e. error(a).
- Form method × replication (B × R) table of yield totals and from this table compute the S.S. due to methods and S.S. due to interaction - Replication × Method i.e. error(b).
- Form spacing × method (A × B) table of yield totals and from this table compute the S.S. due to interaction - Spacing × Method.
- Total S.S. will be obtained as usual by considering all the observations of the experiment and the error S.S. i.e. error(c) will be obtained by subtracting from total S.S. all the S.S. for various sources.
- Now, calculate the mean square for each source of variation by dividing each sum of squares by its respective degrees of freedom.
- Compute the F-value for each source of variation by dividing each mean square by the corresponding error term.
- The analysis of variance table is outlined as follows:

ANOVA

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F**
Replication(R)	(r-1)= 3	ssR	-	-
Spacing(A)	(a-1)= 3	ssA	msA	msA/msE$_1$
Error(a)	(r-1)(a-1)= 9	ssE$_1$	msE$_1$ =E$_a$	
Method(B)	(b-1)= 2	ssB	msB	msB/msE$_2$
Error(b)	(r-1)(b-1)= 6	ssE$_2$	msE$_2$ =E$_b$	
Spacing×Method (A×B)	(a-1)(b-1)= 6	ss(AB)	ms(AB)	ms(AB)/msE$_3$
Error(c)	(r-1)(a-1)(b-1)=18	ssE$_3$	msE$_3$ =E$_c$	
Total	(rab-1)= 47	sstot		

2.5 Standard Errors and Critical Differences

Estimate of S.E. of difference between two A level means = $\sqrt{\frac{2E_a}{rb}}$

Estimate of S.E. of difference between two B level means = $\sqrt{\frac{2E_b}{ra}}$

Estimate of S.E. of difference between two A level means at the same level of B means = $\sqrt{\frac{2\left[(b-1)E_c + E_a\right]}{rb}}$

Estimate of S.E. of difference between two B level means at the same level of A means = $\sqrt{\frac{2\left[(a-1)E_c + E_b\right]}{ra}}$

Critical difference is obtained by multiplying the S.E$_d$ by t$_{5\%}$ table value for respective error d.f. for (i) & (ii). For (iii) & (iv), as the standard error of mean difference involves two error terms, we use the following equation to compute the weighted t values:

$$t = \frac{(b-1)E_c t_c + E_a t_a}{(b-1)E_c + E_a} \text{ , and } t = \frac{(a-1)E_c t_c + E_b t_b}{(a-1)E_c + E_b} \text{ respectively,}$$

where t_a, $t_{b,}$ and t_c are t-values at error d.f. (E_a), error d.f.(E_b) and error d.f.(E_c) respectively.

RESPONSE SURFACE DESIGNS

1. Introduction

The subject of Design of Experiments deals with the statistical methodology needed for making inferences about the treatment effects on the basis of responses (univariate or multivariate) collected through the planned experiments. To deal with the evolution and analysis of methods for probing into mechanism of a system of variables, the experiments involving several factors simultaneously are being conducted in agricultural, horticultural and allied sciences. Data from experiments with levels or level combinations of one or more factors as treatments are normally investigated to compare level effects of the factors and also their interactions. Though such investigations are useful to have objective assessment of the effects of levels actually tried in the experiment, this seems to have inadequate, especially when the factors are quantitative in nature. The above analysis cannot give any information regarding the possible effects of the intervening levels of the factors or their combinations, *i.e.*, one is not able to interpolate the responses at the treatment combinations not tried in the experiment. In such cases, it is more realistic and informative to carry out investigations with the twin purposes:

a) To determine and to quantify the relationship between the response and the settings of a group of experimental factors.
b) To find the settings of the experimental factors that produces the best value or the best set of values of the response(s).

If all the factors are quantitative in nature, it is natural to think the response as a function of the factor levels and data from quantitative factorial experiments can be used to fit the response surfaces over the region of interest. Response surfaces besides inferring about the twin purposes can provide information about the rate of change of a response variable. They can also indicate the interactions between the quantitative treatment factors. The special class of designed experiments for fitting response surfaces is called response surface designs. A good response surface design should possess the properties *viz.*, detectability of lack of fit, the ability to sequentially build up designs of increasing order and the use of a relatively modest, if not minimum, number of design points. Before formulating the problem mathematically, we shall give examples of some experimental situations, where response surface methodology can be usefully employed.

Example 1: The over-use of nitrogen (N) relative to Phosphorus (P) and Potassium (K) concerns both the agronomic and environmental perspective. Phosphatic and Potassic fertilizers have been in short supply and farmers have been more steadily adopting the use of nitrogenous fertilizers because of the impressive virtual response. There is evidence that soil P and K levels are declining. The technique of obtaining individual optimum doses for the N, P and K through separate response curves may also be responsible for unbalanced fertilizer use. Hence, determining the optimum and balanced dose of N, P and K for different crops has been an important issue. This optimum and balanced dose should be recommended to farmers in terms of doses from the different sources and not in terms of the values of N, P and K alone, as the optimum combination may vary from source to source. However, in actual practice the values of N, P and K are given in terms of kg/ha rather than the combined doses alongwith the source of the fertilizers.

Example 2: For value addition to the agriculture produce, food-processing experiments are being conducted. In these experiments, the major objective of the experimenter is to obtain the optimum combination of levels of several factors that are required for the product. To be specific, suppose that an experiment related to osmotic dehydration of the banana slices is to be conducted to obtain the optimum combination of levels of concentration of sugar solution, solution to sample ratio and temperature of osmosis. The levels of the various factors are the following

	Factors	**Levels**
1.	Concentration of sugar solution	40%, 50%, 60%, 70% and 80%
2.	Solution to sample ratio	1:1, 3:1, 5:1, 7:1and 9:1
3.	Temperature of osmosis	25^0C, 35^0C, 45^0C, 55^0C and 65^0C

In this situation, response surface designs for 3 factors each at five equispaced levels can be used.

Example 3: Yardsticks (a measure of the average increase in production per unit input of a given improvement measure) of many fertilizers, manures, irrigation, pesticides for various crops are being obtained and used by planners and administrators in the formulation of policies relating to manufacture/import/subsidy of fertilizers, pesticides, development of irrigation projects etc.

The yardsticks have been obtained from the various factorial experiments. However, these will be more reliable and satisfy more statistical properties, if response surface designs for slope estimation are used.

In general response surface methodology is useful for all the factorial experiments in agricultural experimental programme that are under taken so as to determine the level at which each of these factors must be set in order to optimize the response in some sense and factors are quantitative in nature. To achieve this we postulate that the response is a function of input variables, *i.e.*

$$y_u = \varphi(x_{1u}, x_{2u}, ..., x_{vu}) + e_u \qquad (1.1)$$

where $u = 1,2,..., N$ represents the N observations and x_{iu} is the level of the i^{th} factor in the u^{th} observation. The function φ describes the form in which the response and the input variables are related and e_u is the experimental error associated with the u^{th}observation such that E (e_u) = 0 and Var(e_u) = σ^2. Knowledge of function φgives a complete summary of the results of the experiment and also enables us to predict the response for values of the x_{iu} that are not included in the experiment. If the function $\varphi$is known then using methods of calculus, one may obtain the values of $x_1, x_2, ..., x_v$ which give the optimum (say, maximum) response. In practice the mathematical form of $\varphi$is not known; we, therefore, often approximate it, within the experimental region, by a polynomial of suitable degree in variables x_{iu}. The adequacy of the fitted polynomial is tested through the usual analysis of variance. Polynomials which adequately represent the true dose-response relationship are called **Response Surfaces** and the designs that allow the fitting of response surfaces and provide a measure for testing their adequacy are called **response surface designs**. If the function φin (1.1) is of degree one in $x_{iu}'s$ *i.e.*

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + ... + \beta_v x_{vu} + e_u \qquad (1.2)$$

we call it a first-order response surface in $x_1, x_2, ..., x_v$. If (1.1) takes the form

$$y_u = \beta_0 + \sum_{i=1}^{v}\beta_i x_{iu} + \sum_{i=1}^{v}\beta_{ii} x_{iu}^2 + \sum_{i=1}^{v-1}\sum_{i'=i+1}^{v}\beta_{ii'} x_{iu} x_{i'u} + e_u \qquad (1.3)$$

We call it a second-order (quadratic) response surface. Henceforth, we shall concentrate on the second order response surface which is more useful in agricultural experiments.

2. The Quadratic Response Surface

The general form of a second-degree (quadratic) surface is

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + ... + \beta_v x_{vu} + \beta_{11} x_{1u}^2 + \beta_{22} x_{2u}^2 + ... + \beta_{vv} x_{vu}^2 +$$
$$\beta_{12} x_{1u} x_{2u} + \beta_{13} x_{1u} x_{3u} + ... + \beta_{v-1,v} x_{v-1,u} x_{vu} + e_u$$

Let us assume that $x_{iu}'s$ satisfy the following conditions:

(A) $\sum_{u=1}^{N}\left\{\prod_{u=1}^{v} x_{iu}^{\alpha_i}\right\} = 0$, if any α_i is odd, for $\alpha_i = 0,1,2 \; or 3$ and $\sum \alpha_i \le 4$.

(B) $\sum_{u=1}^{N} x_{iu}^2 =$ constant (for all i) $= N\lambda_2$ (say)

(C) $\sum_{u=1}^{N} x_{iu}^4 =$ constant (for all i) = $CN\lambda_4$ (say) (2.1)

(D) $\sum_{u=1}^{N} x_{iu}^2 x_{i'u}^2 =$ constant $= N\lambda_4$ (say), for all $i \ne i'$

We shall estimate the parameters $\beta_i's$ through the method of least squares. Let $b_0, b_i's, b_{ii}'s, b_{ii'}'s$ denote the best linear unbiased estimate of $\beta_0, \beta_i's, \beta_{ii}'s, \beta_{ii'}'s$ respectively. Under the above restrictions on $x_{iu}'s$, the normal equations are found to be:

$$\sum_{u=1}^{N} y_u = Nb_0 + N\lambda_2 \sum_{i=1}^{v} b_{ii}$$
$$\sum_{u=1}^{N} x_{iu} y_u = N\lambda_2 b_i$$
$$\sum_{u=1}^{N} x_{iu} x_{i'u} y_u = N\lambda_4 b_{ii'} \qquad (2.2)$$
$$\sum_{u=1}^{N} x_{iu}^2 y_u = N\lambda_2 b_0 + CN\lambda_4 b_{ii} + N\lambda_4 \sum_{i' \ne i} b_{i'i'}$$
$$= N\lambda_2 b_0 + (C-1)N\lambda_4 b_{ii} + N\lambda_4 \sum_{i=1}^{v} b_{ii}$$

Solving the above normal equations, we obtain the estimates b_i*'s* as

$$
\begin{aligned}
b_0 &= \left[\lambda_4(C+v-1)\sum_{u=1}^{N} y_u - \lambda_2 \sum_{i=1}^{v}\sum_{u=1}^{N} x_{iu}^2 y_u\right] \Big/ N\Delta \\
b_i &= \sum_{u=1}^{N} x_{iu} y_u \Big/ N\lambda_2 \\
b_{ii'} &= \sum_{u=1}^{N} x_{iu} x_{i'u} y_u \Big/ N\lambda_4 \\
b_{ii} &= \left[\sum_{u=1}^{N} x_{iu}^2 y_u - \left\{\left(\lambda_2^2 - \lambda_4\right)\sum_{i=1}^{v}\sum_{u=1}^{N} x_{iu}^2 y_u - (C-1)\lambda_2\lambda_4 \sum_{u=1}^{N} y_u\right\} \Big/ \Delta\right] \Big/ [(C-1)N\lambda_4]
\end{aligned}
\tag{2.3}
$$

where $\Delta = (C+v-1)\lambda_4 - v\lambda_2^2$.

The variances of and covariances between the estimated parameters are as follows:

$$
\begin{aligned}
V(b_0) &= \lambda_4(C+v-1)\sigma^2 \big/ N\Delta \\
V(b_i) &= \sigma^2 \big/ N\lambda_2 \\
V(b_{ii'}) &= \sigma^2 \big/ N\lambda_4 \\
V(b_{ii}) &= \sigma^2\left[1+\left(\lambda_2^2 - \lambda_4\right)\right] \big/ [(C-1)N\lambda_4] \\
Cov(b_0, b_{ii}) &= -\lambda_2 \sigma^2 \big/ N\Delta \\
Cov(b_{ii}, b_{i'i'}) &= \left(\lambda_2^2 - \lambda_4\right)\sigma^2 \big/ [((C-1)N\lambda_4\Delta]
\end{aligned}
\tag{2.4}
$$

Other covariances are zero. From the above expressions it is clear that a necessary condition for the design to exist is that $\Delta > 0$. Thus, a necessary condition for a Second Order Design to exist is that

(E) $\lambda_4 \big/ \lambda_2^2 > v/(C+v-1)$ (2.5)

If $\hat{y}$ is the estimated response at any given experimental point $(x_{10}, x_{20}, ..., x_{v0})$, then the variance of $\hat{y}$ is given by

$$
\begin{aligned}
V(\hat{y}) = V(b_0) + V(b_i)\left(\sum_{i=1}^{v} x_{i0}^2\right) + V(b_{ii})\left(\sum_{i=1}^{v} x_{i0}^4\right) + V(b_{ii'})\left(\sum_{i=1}^{v-1}\sum_{i'=i+1}^{v} x_{i0}^2 x_{i'0}^2\right) \\
+ 2Cov(b_0, b_{ii})\left(\sum_{i=1}^{v} x_{i0}^2\right) + 2Cov(b_{ii}, b_{i'i'})\left(\sum_{i=1}^{v-1}\sum_{i'=1+1}^{v} x_{i0}^2 x_{i'0}^2\right)
\end{aligned}
\tag{2.6}
$$

If $\sum_{i=1}^{v} x_{i0}^2 = d^2$, where d is the distance of the point $(x_{10}, x_{20}, ..., x_{v0})$ from the origin, then we may write

$$
\begin{aligned}
V(\hat{y}) = V(b_0) + d^2[V(b_i) + 2Cov(b_0, b_{ii})] + d^4 \mathrm{V}(b_{ii}) \\
+ \sum_{i=1}^{v-1}\sum_{i'=i+1}^{v} x_{i0}^2 x_{i'0}^2 [V(b_{ii'}) + 2Cov(b_{ii}, b_{i'i'}) - 2V(b_{ii})]
\end{aligned}
\tag{2.7}
$$

From the above expression, it is clear that if the coefficient of $\sum_{i=1}^{v-1}\sum_{i'=i+1}^{v} x_{i0}^2 x_{i'0}^2$ is made equal to zero, the variance of the estimated response at $(x_{10}, x_{20}, ..., x_{v0})$ will be a function of d, the distance of the point $(x_{10}, x_{20}, ..., x_{v0})$ from the origin. Now, the coefficient of $\sum_{i=1}^{v-1}\sum_{i'=i+1}^{v} x_{i0}^2 x_{i'0}^2$ is

$$V(b_{ii'}) + 2\ Cov(b_{ii}, b_{i'i'}) - 2V(b_{ii})$$

$$= \frac{\sigma^2}{N\lambda_4}\left[1 + \frac{2(\lambda_2^2 - \lambda_4)}{\varDelta(C-1)} - \frac{2}{(C-1)}\left\{1 + \frac{\lambda_2^2 - \lambda_4}{\varDelta}\right\}\right] \tag{2.8}$$

$$= \frac{\sigma^2}{N\lambda_4}\left[1 - \frac{2}{(C-1)}\right]$$

Obviously, this is zero, if and only if $C = 3$. Thus, when $C = 3$, the variance of the estimated response at a given point, the response being estimated through a design satisfying (A), (B), (C), (D), (E) becomes a function of the distance of that point from the origin. Such designs are called as Second Order Rotatable Designs (SORD). We may now formally define a SORD:

Let us consider N treatment combinations (points) $\{x_{iu}\}$, $i = 1,2,...,v, u = 1,2,...,N$ to form a design in v factors, through which a Second-degree surface can be fitted. This design is said to be a SORD if the variance of the estimated response at any given point is a function of the distance of that point from the origin. The necessary and sufficient conditions for a set of points $\{x_{iu}\}$, $i = 1,2,...,v, u = 1,2,...,N$ to form a SORD are

(A') $\sum_{u=1}^{N}\left\{\prod_{u=1}^{v} x_{iu}^{\alpha_i}\right\} = 0$, if any α_i is odd, for $\alpha_i = 0,1,2\ or 3$ and $\sum \alpha_i \le 4$.

(B') $\sum_{u=1}^{N} x_{iu}^2 = N\lambda_2$

(C') $\sum_u x_{iu}^4$ = constant = $3\ N\lambda_4$ $\qquad i = 1,2,...,v$

(D') $\sum_u x_{iu}^2 x_{i'u}^2 = N\lambda_4$; $\qquad i \ne i'$

(E') $\lambda_4 / \lambda_2^2 > v/(v+2)$ (2.9)

The conditions (A'), (B') and (D') are same as conditions (A), (B) and (D) in (2.1).

We now prove the following.

Lemma: If a set of points $\{x_{iu},\ i = 1,2,...,v, u = 1,2,...,N\}$, satisfying (A'), (B'), (C') and (D') are such that every point is equidistant from the origin, then

$$\lambda_4 / \lambda_2^2 = v/(v+2) \tag{2.10}$$

Proof: Let d be the distance of any point from the origin. Then, since all the points are equidistant from the origin, we have

$$d^2 = \frac{1}{N}\sum_{u=1}^{N}\left(\sum_{i=1}^{v} x_{iu}^2\right) = v\lambda_2$$

$$d^4 = \frac{1}{N}\sum_{u=1}^{N}\left(\sum_{i=1}^{v} x_{iu}^2\right)^2$$

and
$$= \frac{1}{N}\sum_{u=1}^{N}\left[\sum_{i=1}^{v} x_{iu}^4 + 2\sum_{i=1}^{v}\sum_{i'=i+1}^{v} x_{iu}^2 x_{i'u}^2\right]$$

$$= 3v\lambda_4 + v(v-1)\lambda_4$$

Thus, $v^2\lambda_2^2 = 3v\lambda_4 + v(v-1)\lambda_4$

or, $\lambda_4(v+2) - v\lambda_2^2 = 0$

An arrangement of points satisfying (A’), (B’), (C’) and (D’) but not (E’) is called a Second Order Rotatable Arrangement (SORA). A SORA can always be converted to an SORD by adding at least one central point.

A near stationary region is defined as a region where the surface slopes along the v variable axes are small compared to the estimate of experimental error. The stationary point of a near stationary region is the point at which the slope of the response surface is zero when taken in all the directions. The coordinates of the stationary point $\mathbf{x_0} = (x_{10}, x_{20}, ..., x_{v0})'$ are obtained by differentiating the following estimated response equation with respect to each x_i and equating the derivatives to zero and solving the resulting equations

$$\hat{Y}(x) = b_0 + \sum_{i=1}^{v} b_i x_i + \sum_{i=1}^{v} b_{ii} x_i^2 + \sum_{i=1}^{v-1}\sum_{i'=i+1}^{v} b_{ii'} x_i x_{i'} \qquad (2.11)$$

In matrix notation (2.11) can be written as

$$\hat{Y}(x) = b_0 + \mathbf{x'b} + \mathbf{x'Bx} \qquad (2.12)$$

where $\mathbf{x} = (x_1, x_2, ..., x_v)'$, $\mathbf{b} = (b_1, b_2, ..., b_v)'$

and

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12}/2 & \dots & b_{1v}/2 \\ b_{12}/2 & b_{22} & \dots & b_{2v}/2 \\ \dots & \dots & \dots & \dots \\ b_{1v}/2 & b_{2v}/2 & \dots & b_{vv} \end{bmatrix}.$$

From equation (2.12)

$$\frac{\partial \hat{Y}(x)}{\partial x} = \mathbf{b} + 2\mathbf{Bx} \qquad (2.13)$$

The stationary point $\mathbf{x_0}$ is obtained by equating (2.13) to zero and solving for $\mathbf{x}$, *i.e.*

$$\mathbf{x_0} = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} \tag{2.14}$$

To find the nature of the surface at the stationary point we examine the second derivative of $\hat{Y}(x)$. From (2.13)

$$\frac{\partial^2 \hat{Y}(x)}{\partial x^2} = 2\mathbf{B} \qquad \text{(since } \mathbf{B} \text{ is symmetric).}$$

The stationary point is a maximum, minimum or a saddle point according as **B** is negative definite, positive definite or indefinite matrix. If $\lambda_1, \lambda_2, ..., \lambda_v$ represent the *v* eigenvalues of **B**. Then it is easy to see that if $\lambda_1, \lambda_2, ..., \lambda_v$ are

(i) All negative, then at $\mathbf{x}_0$ the surface is a maximum

(ii) All positive, then at $\mathbf{x}_0$ the surface is a minimum

(iii) of mixed signs, i.e. some are positive and others are negative, then x_0 is a saddle point of the fitted surface.

Furthermore, if λ_i is zero (or very close to zero), then the response does not change in value in the direction of the axis associated with x_i variable. The magnitude of λ_i indicates how quickly the response changes in the direction of axis associated with x_i variable.

The conditions in (2.1) and (2.9) help in fitting of the response surfaces and define some statistical properties of the design like rotatability. However, these conditions need not necessary be satisfied before fitting a response surface. This can be achieved by using the software packages like the Statistical Analysis System (SAS). PROC RSREG fits a second order response surface design and locates the coordinates of the stationary point, predict the response at the stationary point and give the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_v$ and the corresponding eigen vectors. It also helps in determining whether the stationery point is a point of maxima, minima or is a saddle point. The lack of fit of a second order response surface can also be tested using LACKFIT option under model statement in PROC RSREG. The lack of fit is tested using the statistic

$$F = \frac{SS_{LOF} / (N'\text{-}p)}{SS_{PE} / (N - N')} \tag{2.15}$$

where *N* is the total number of observations, *N'* is the number of distinct treatments and *p* is the number of terms included in the model. SS_{PE} (sum of squares due to pure error) has been calculated in the following manner: denote the l^{th} observation at the u^{th} design point by y_{lu}, where $l = 1, \ldots, r_u (\geq 1), u = 1, \ldots, N'$. Define $\bar{y}_u$ to be average of r_u observations at the u^{th} design point. Then, the sum of squares for pure error is

$$SS_{PE} = \sum_{u=1}^{N'} \sum_{l=1}^{r_u} (y_{lu} - \bar{y}_u)^2 \tag{2.16}$$

Then sum of squares due to lack of fit (SS_{LOF}) = sum of squares due to error - SS_{PE}

The analysis of variance table for a second order response surface design is given below.

Table 1. Analysis of Variance for second order response surface

Source	d.f.	S.S.
Due to regression coefficients	$2v+\binom{v}{2}$	$\hat{b}_0\sum_{u=1}^{N} y_u + \sum_{i}\hat{b}_i\left(\sum_{u=1}^{N} x_{iu}y_u\right) + \sum_{i}\hat{b}_{ii}\left(\sum_{u=1}^{N} x_{iu}^2 y_u\right)$ $+\sum\sum_{i\neq i'}\hat{b}_{ii'}\left(\sum_{u=1}^{N} x_{iu}x_{i'u}y_u\right) - CF$
Error	$N-2v-\binom{v}{2}-1$	*By subtraction = SSE*
Total	$N-1$	$\sum_{u=1}^{N} y_u^2 - CF$

In the above table CF = correction factor = $\frac{(Grand Total)^2}{N}$. For testing the lack of fit the sum of squares is obtained using (2.16) and then sum of squares is obtained by subtracting the sum of squares due to pure error from sum of squares due to error. The sum of squares due to lack of fit and sum of squares due to pure error are based on $N'-2v-\binom{v}{2}-1$ and $N-N'$ degrees of freedom respectively.

It is suggested that in the experiments conducted to find a optimum combination of levels of several quantitative input factors, at least one level of each of the factors should be higher than the expected optimum. It is also suggested that the optimum combination should be determined from response surface fitting rather than response curve fitting, if the experiment involves two or more than two factors.

3. Construction of Second Order Rotatable designs

A second order response surface design is at least resolution V fractional factorial design. Here

3.1 Central Composite Rotatable Designs

Let there be v factors in the design. A class of SORD for v factors can be constructed in the following manner. Construct a factorial v-factors with levels $\pm\alpha$ containing 2^p combinations, where 2^p is the smallest fraction of 2^v without confounding any interaction of third order or less. Next, another $2v$ points of the following type are considered: $(\pm\beta\ 0\ 0\ \ldots\ 0)$, $(0\ \pm\beta\ 0\ \ldots\ 0)$, $(\ 0\ 0\ \ldots\ \pm\beta)$. These $N=2^p+2v$ points, give rise to a SORD in v factors with levels $\pm\alpha,\ \pm\beta,\ 0$. For this design,

$$\sum_{u=1}^{N} x_{iu}^2 = 2^p\alpha^2 + 2\beta^2$$

$$\sum_{u=1}^{N} x_{iu}^4 = 2^p\alpha^4 + 2\beta^4$$

$$\sum_{u=1}^{N} x_{i0}^2 x_{i'0}^2 = 2^p\alpha^4.$$

On applying the condition of rotatability,

$$\begin{aligned}&3.2^{p}\alpha^{4} = 2^{p}\alpha^{4} + 2\beta^{4}\\&\Rightarrow \beta^{4} = \alpha^{4}2^{p}\\&or\ \beta^{2}\big/\alpha^{2} = 2^{p/2}.\end{aligned}$$

This equation gives a relationship between β and α. For determining α and β uniquely, we either fix $\alpha = 1$ or $\lambda_2 = 1$. For $\alpha = 1, \Rightarrow \beta^2 = 2^{p/2}$.

Example. Let $v = 4$. Then the points of the SORD are

$-\alpha$	$-\alpha$	$-\alpha$	$-\alpha$
$-\alpha$	$-\alpha$	$-\alpha$	α
$-\alpha$	$-\alpha$	α	$-\alpha$
$-\alpha$	$-\alpha$	α	α
$-\alpha$	α	$-\alpha$	$-\alpha$
$-\alpha$	α	$-\alpha$	α
$-\alpha$	α	α	$-\alpha$
$-\alpha$	α	α	α
α	$-\alpha$	$-\alpha$	$-\alpha$
α	$-\alpha$	$-\alpha$	α
α	$-\alpha$	α	$-\alpha$
α	$-\alpha$	α	α
α	α	$-\alpha$	$-\alpha$
α	α	$-\alpha$	α
α	α	α	$-\alpha$
α	α	α	α
β	0	0	0
$-\beta$	0	0	0
0	β	0	0
0	$-\beta$	0	0
0	0	β	0
0	0	$-\beta$	0
0	0	0	β
0	0	0	$-\beta$
0	0	0	0

There are *25* points – a central point has been added because, all the non-central points are equidistant from the origin, as $\beta = 2\alpha$, here.

3.2 Construction of SORD using BIB Designs

If there exists a BIB design D with parameters $v^*, b^*, r^*, k^*, \lambda^*$ such that $r^* = 3\lambda^*$, then a SORD with each factor at *3* levels can be constructed.

Let $\mathbf{N}^*$be the $v^* \times b^*$ incidence matrix of D. Then $\mathbf{N}^{*\prime}$ is a matrix of order $b^* \times v^*$, every row of which contains exactly k^* unities and every column contains exactly r^*unities, rest positions being filled up by zeros. In $\mathbf{N}^{*\prime}$, replace the unity by α. Then, we get b^*combinations involving α and zero. Next, each of these combinations are 'multiplied' with those of a

2^{k^*}factorial with levels ± 1 where, the term 'multiplication' means the multiplication of the corresponding entries in the two combinations, zero entries remaining unaltered. Thus, if $(\alpha\ \alpha\ 0)$ is multiplied by $(-1\ -1)$ we get $(-\alpha\ \ -\alpha\ \ 0)$. The procedure of multiplication gives rise to $b^*\,2^{k^*}$ points each of v^*-dimension. These points evidently satisfy all the conditions (A'), (B'), (C') and (D'); however, since each point in the arrangement is at the same distance from the origin, we have to take at least one central point to get a SORD in $v = v^*$factors. The levels of the factors are $\pm\alpha, 0$. The value of α can be determined by fixing $\lambda_2 = 1$.

SORD's can be constructed using BIB designs, even when $r^* \neq 3\lambda^*$. In the case, where $r^* < 3\lambda^*$ the set of $b^*\,2^{k^*}$points obtained using $\mathbf{N}^{*\prime}$ is to be augmented with further $2v^*$points of the type

$$(\pm\beta\ 0\ 0\ \ldots\ 0),\ (0\ \pm\beta\ 0\ \ldots\ 0),\ (0\ 0\ \ldots\ \pm\beta)$$

For the N points ($N = b^*\,2^{k^*} +\ 2v^*$), we have

$$\sum_u x_{iu}^4 = r^* 2^{k^*}\alpha^4 + 2\beta^4$$

$$\sum_u x_{iu}^2 x_{i'u}^2 = \lambda^* 2^{k^*}\alpha^4.$$

Thus $2\beta^4 + r^* 2^{k^*}\alpha^4 = 3\lambda^* 2^{k^*}\alpha^4$

or, $\beta^2/\alpha^2 = (3\lambda^* - r^*)^{1/2}.2^{\frac{k^*-1}{2}}$

When $r^* > 3\lambda^*$, the points augmented are of type $(\pm\beta\ \pm\beta\ \ldots\ldots \pm\beta)$ and 2^p in number, where 2^p is the smallest fraction of 2^{v^*} factorial with levels $\pm\beta$, such that no interaction of order three or less is confounded. In this case,

$$\sum_u x_{iu}^4 = r^* 2^{k^*}\alpha^4 + 2^p\beta^4$$

$$\sum_u x_{iu}^2 x_{i'u}^2 = \lambda^* 2^{k^*}\alpha^4 + 2^p\beta^4.$$

Thus, $3\lambda^* 2^{k^*}\alpha^4 + 3.2^p\beta^4 = r^* 2^{k^*}\alpha^4 + 2^p\beta^4$

or, $2^{p+1}\beta^4 = (r^* - 3\lambda^*)2^{k^*}\alpha^4$,

which gives $\beta^2/\alpha^2 = (r^* - 3\lambda^*)^{1/2}.2^{(k^*-p-1)/2}$.

In both the cases, we get v^*-factor SORD with each factor at five levels

4. Practical Exercise

Exercise 1: Consider an experiment that was conducted to investigate the effects of three fertilizer ingredients on the yield of a crop under fields conditions using a second order rotatable design. The fertilizer ingredients and actual amount applied were nitrogen (N), from 0.89 to 2.83 kg/plot; phosphoric acid (P_2O_5) from 0.265 to 1.336 kg/plot; and potash (K_2O), from 0.27 to 1.89 kg/plot. The response of interest is the average yield in kg per plot. The levels of nitrogen, phosphoric acid and potash are coded, and the coded variables are defined as

X_1=(N-1.629)/0.716, X_2=(P_2O_5-0.796)/0.311, X_3=(K_2O -1.089)/0.482

The values 1.629, 0.796 and 1.089 kg/plot represent the centres of the values for nitrogen, phosphoric acid and potash, respectively. Five levels of each variable are used in the experimental design. The coded and measured levels for the variables are listed as

	Levels of x_I				
	-1.682	**-1.000**	**0.000**	**+1.000**	**+1.682**
N	0.425	0.913	1.629	2.345	2.833
P_2O_5	0.266	0.481	0.796	1.111	1.326
K_2O	0.278	0.607	1.089	1.571	1.899

Six center point replications were run in order to obtain an estimate of the experimental error variance. The complete second order model to be fitted to yield values is

$$Y=\beta_0+\sum_{i=1}^{3}\beta_i x_i+\sum_{i=1}^{3}\beta_{ii}x_i^2+\sum_{i=1}^{2}\sum_{i'=2}^{3}\beta_{ii'}x_i x_{i'}+e$$

The following table list the design settings of x_1, x_2 and x_3 and the observed values at 15 design points N, P_2O_5, K_2O and yield are in kg.

Table 2: Central Composite Rotatable Design Settings in the Coded Variables x_1, x_2 and x_3, the original variables N, P_2O_5, K_2O and the Average Yield of a Crop at Each Setting

x_1	x_2	x_3	**N**	**P_2O_5**	**K_2O**	**Yield**
-1	-1	-1	0.913	0.481	0.607	5.076
1	-1	-1	2.345	0.481	0.607	3.798
-1	1	-1	0.913	1.111	0.607	3.798
1	1	-1	2.345	1.111	0.607	3.469
-1	-1	1	0.913	0.481	1.571	4.023
1	-1	1	2.345	0.481	1.571	4.905
-1	1	1	0.913	1.111	1.571	5.287
1	1	1	2.345	1.111	1.571	4.963
-1.682	0	0	0.425	0.796	1.089	3.541
1.682	0	0	2.833	0.796	1.089	3.541
0	-1.682	0	1.629	0.266	1.089	5.436
0	1.682	0	1.629	1.326	1.089	4.977
0	0	-1.682	1.629	0.796	0.278	3.591
0	0	1.682	1.629	0.796	1.899	4.693
0	0	0	1.629	0.796	1.089	4.563
0	0	0	1.629	0.796	1.089	4.599
0	0	0	1.629	0.796	1.089	4.599
0	0	0	1.629	0.796	1.089	4.275
0	0	0	1.629	0.796	1.089	5.188
0	0	0	1.629	0.796	1.089	4.959

The output for the above problem is as follows:

Response Surface for Variable YIELD

Response Mean	4.464050
Root MSE	0.356424
R-Square	0.8440
Coef. of Variation	7.9843

Regression	**d.f.**	**Sum of Squares**	**R-Square**	**F-Ratio**	**Prob > F**
Linear	3	1.914067	0.2350	5.022	0.0223
Quadratic	3	3.293541	0.4044	8.642	0.0040
Crossproduct	3	1.666539	0.2046	4.373	0.0327
Total Regression	9	6.874147	0.8440	6.012	0.0049

Regression	**d.f.**	**Sum of Squares**	**R-Square**	**F-Ratio**	**Prob > F**
Lack of Fit	5	0.745407	0.149081	1.420	0.3549
Pure Error	5	0.524973	0.104995		
Total Error	10	1.270380	0.127038		

Parameter	**d.f**	**Estimate**	**Std Error**	**T-ratio**	**Prob > |T|**
INTERCEPT	1	6.084180	1.543975	3.941	0.0028
N	1	1.558870	0.854546	1.824	0.0981
P	1	-6.009301	2.001253	-3.003	0.0133
K	1	-0.897830	1.266909	-0.709	0.4947
N*N	1	-0.738716	0.183184	-4.033	0.0024
P*N	1	-0.142436	0.558725	-0.255	0.8039
P*P	1	2.116594	0.945550	2.238	0.0491
K*N	1	0.784166	0.365142	2.148	0.0573
K*P	1	2.411414	0.829973	2.905	0.0157
K*K	1	-0.714584	0.404233	-1.768	0.1075

Factor	**d.f.**	**Sum of Squares**	**Mean Squares**	**F-Ratio**	**Prob > F**
N	4	2.740664	0.685166	5.393	0.0141
P	4	1.799019	0.449755	3.540	0.0477
K	4	3.807069	0.951767	7.492	0.0047

Canonical Analysis of Response Surface

Factor	**Critical Value**
N	1.758160
P	0.656278
K	1.443790

Predicted value at stationary point 4.834526 kg

	Eigenvectors		
Eigenvalues	**N**	**P**	**K**
2.561918	0.021051	0.937448	0.347487
-0.504592	0.857206	-0.195800	0.476298
-1.394032	-0.514543	-0.287842	0.807708

Stationary point is a saddle point.

The eigenvalues obtained are λ_1, λ_2 and λ_3 as 2.561918, -0.504592, -1.394032. As λ_2 and λ_3 are negative, therefore, take $W_2 = W_3 = 0$. Let

M = {0.021051 0.857206 -0.514543,
0.937448 -0.195800 -0.287842,
0.34787 0.476298 0.807708};

denotes the matrix of eigenvectors. The estimated response at the stationary points be 4.834526 kg/plot. Let the desired response be Y_{des}=5.0 kg/plot. Therefore, let W_1, obtained from the equation is sqrt (difference/2.561918)=AX1, say. To obtain various different sets of many values of W_1, generate a random variable, u, which follows uniform distribution and multiply this value with $2u-1$ such that W_1 lies within the interval, (-AX1, AX1). Now to get a combination of $x_i's$ that produces the desired response obtain $\mathbf{x} = \mathbf{M} * \mathbf{W} + \mathbf{x_0}$.

Combinations of N, P, K estimated to produce 5.0 kg/plot of Beans.

Y	N	P	K
5.0	1.760	0.730	1.471
	1.762	0.815	1.503
	1.754	0.460	1.371

One can select a practically feasible combination of N, P and K.

5. Response Surface Designs for Slope Estimation

The above discussion relates to the response surface designs for response optimization. In many practical situations, however, the experimenter is interested in estimation of the rate of change of response for given value of independent variable(s) rather than optimization of response. This problem is frequently encountered *e.g.*, in estimating rates of reaction in chemical experiments; rates of growth of biological populations; rates of changes in response of a human being or an animal to a drug dosage, rate of change of yield per unit of fertilizer dose. Efforts have been made in the literature for obtaining efficient designs for the estimation of differences in responses *i.e.*, for estimating the slope of a response surface.

Many researchers with different approaches have taken up the problem of designs for estimating the slope of a response surface. We confine ourselves to two main approaches, namely

- Slope Rotatability
- Minimax Designs

The designs possessing the property that the estimate of derivative is equal for all points equidistant from the origin are known as **slope rotatable designs**. For a second order response surface, the rate of change of response due to i^{th} independent variable is given by

$$\frac{\partial \hat{y}(x)}{\partial x_i} = b_i + 2b_{ii}x_i + \sum_{i' \neq i}^{v} b_{ii'}x_{i'}$$

For second order design satisfying (2.1) we have

$$Cov(b_i, b_{ii}) = Cov(b_i, b_{ij}) = Cov(b_{ii}, b_{ii'})$$

Thus variance of $\frac{\partial \hat{y}(x)}{\partial x_i}$ is given by

$$Var\left(\frac{\partial\hat{y}(x)}{\partial x_i}\right)=Var(b_i)+\rho^2 Var(b_{ii})+x_i^2\left[4Var(b_{ii})-Var(b_{ii'})\right]$$

Thus in order to obtain slope rotatable design, the design must satisfy

- Conditions of symmetry (2.1)
- $\frac{\lambda_4}{\lambda_2^2} > v/(c+v-1)$
- $4Var(b_{ii})=Var(b_{ii'})$.

It is important to note here that no rotatable design can be slope rotatable.

A minimax design is one that minimizes the variance of the estimated slope maximized over all points in the design.

DESIGNS FOR MIXTURE EXPERIMENTS

Mixtures are formed by blending or mixing two or more components. Some common examples are:

(a) Construction concrete (mixture of sand, water and cement)
(b) Railroad flares (product of blending together Magnesium, Sodium Nitrate, Strontium Nitrate and binder)
(c) Fruit punch (mixture of juices of watermelon, orange and pineapple)
(d) Fertilizer mixture (mixture of Potash, Rock Phosphate, Super Phosphate and urea)
(e) Cake formulation (blend of baking powder, shortening, flour, sugar, and water)

The manufacturers of such products are interested in one or more properties of the final product. For example, in construction concrete, the hardness or compression strength of the mixture is of interest; in railway flares, the illumination and duration of the illumination of the flares are the interesting properties; in fruit punch the fruitiness flavor of the punch is the property of interest: in fertilizer mixture, the crop yield is of interest; and in cake formulation, the property of interest is the fluffiness of the cake or the layered appearance. The property of the final product depends on the percentage or proportions of the ingredients mixed.

Another reason for mixing two or more ingredients is to see whether the blend has a more desirable property than the individual ingredients. For example, suppose there are three types of gasoline, A, B and C, in stock. One may be interested in the antiknock rating of the stocks, used singly and in combination. That is, one may want to know if there exists some combination of the three which yields higher antiknock rating than the three used singly. If that is true, naturally one would go for the combination rather than any for the single stocks!

As the property of the final product depends on the mixture combination, one would be interested to study the functional relationship between the measured property or measured response and the mixing proportions of the ingredients. From experimental viewpoint, such a study is of interest in order (i) to determine some combination of the mixture ingredients that would be best in some sense, or (ii) to have a better understanding of the effects of the ingredients on the response.

Consider a mixing blend with q components in the proportions $\boldsymbol{x} = (x_1, x_2, ..., x_q)$ .Clearly, 0≤ x_i≤ 1, and 1 ≤ i ≤ q, and

$$\sum_{i=1}^{q} x_i = 1. \qquad (1)$$

Because of the natural constraint (1), a mixture experiment belongs to a class of its own. It has vast application in different research areas and also industries, such as

- Agriculture
- Engineering
- Pharmaceutical
- Biomedical
- Horticulture

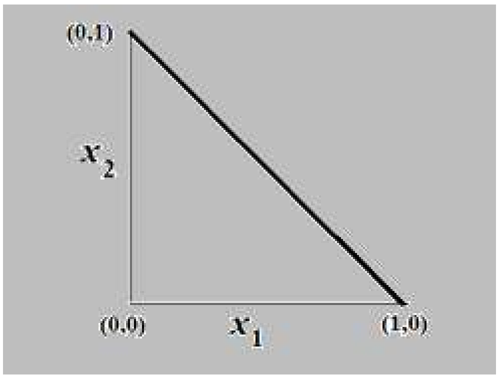
etc

The experimental region for a mixture experiment is given by

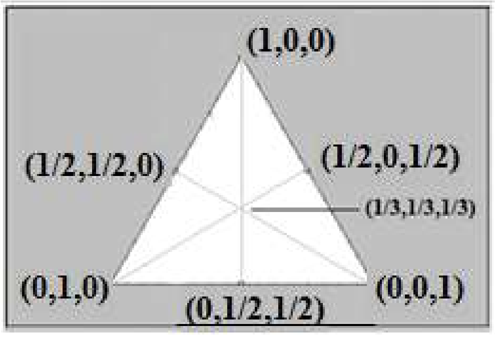
$$\Xi = \{(x_1, x_2, ..., x_q) : x_i \geq 0,\ 1 \leq i \leq q,\ \sum_{i=1}^{q} x_i = 1\}$$

Geometrically, Ξ is represented by a q-1 dimensional simplex. The vertex points of the simplex region are of the type (1,0,...,0), (0,1,0,...,0), ..., (0,0,...,0,1).These points are called **pure or single point mixtures.** The experimental points lie within or on the boundary of the simplex region.

For example, consider a mixture of two components with mixing proportions $(x_1, x_2), 0 \le x_1, x_2 \le 1, x_1 + x_2 = 1.$ Here the experimental region will be a straight line with the end point (0, 1) and (1, 0):



For a mixture of three components with mixing proportions $(x_1, x_2, x_3), 0 \le x_1, x_2, x_3 \le 1, x_1 + x_2 + x_3 = 1,$ the experimental region is a triangle with vertex points (1,0,0), (0,1,0) and (0,0,1):



The points (1/2, 1/2,0), (1/2,0, 1/2) and (0, 1/2, 1/2) are called the **mid-points of the edges**, and the point (1/3,1/3,1/3) is the overall **centroid point**.

For a 4-component mixture, the experimental region is a tetrahedron with four extreme points and six mid-points of edges.

Let $Y_{\boldsymbol{x}}$ denote the response corresponding to the mixture combination $\boldsymbol{x}$. Scheffé (1958) first defined models for expressing the response in terms of the mixing proportions of the ingredients. The models are as follows:

Linear (homogeneous): $$Y_{\boldsymbol{x}} = \sum_{i=1}^{q} \beta_i x_i + \varepsilon$$

Quadratic : $$Y_{\boldsymbol{x}} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j=1}^{q} \beta_{ij} x_i x_j + \varepsilon$$

Full cubic : $$Y_{\boldsymbol{x}} = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j=1}^{q} \beta_{ij} x_i x_j + \sum_{i<j=1}^{q} \delta_{ij} x_i x_j (x_i - x_j) + \sum_{i<j<k=1}^{q} \beta_{ijk} x_i x_j x_k + \varepsilon$$

Special cubic : $$Y_{\mathbf{x}} = \sum_{i=1}^{q}\beta_i x_i + \sum_{i<j=1}^{q}\beta_{ij}x_i x_j + \sum_{i<j<k=1}^{q}\beta_{ijk}x_i x_j x_k + \varepsilon,$$

where ε is the error term assumed to be distributed with mean zero and variance σ^2.

The quadratic model is found to be appropriate in most situations.

A mixture experiment is conducted to estimate the parameters of the fitted model or to estimate some functions of the model parameters, like say the optimum mixture combination that optimizes the expected response. There are several different types of designs for a mixture experiment. The most common ones are the simplex lattice and the simplex centroid designs. Other common designs are the simplex axial and extreme vertex designs. Each design is used for a different purpose as listed below:

- If there are many components in a mixture, the first choice is to screen out the most important ones. The simplex axial and simplex centroid designs are used for this purpose.
- If the number of components is not large, but a high order polynomial equation is needed in order to accurately describe the response, then a simplex lattice design can be used.
- Extreme vertex designs are used for the cases when there are constraints on one or more components (e.g., if the proportion of watermelon juice in a fruit punch recipe is required to be less than 30%, and the combined proportion of watermelon and orange juice should always be between 40% and 70%).

Simplex Lattice Design

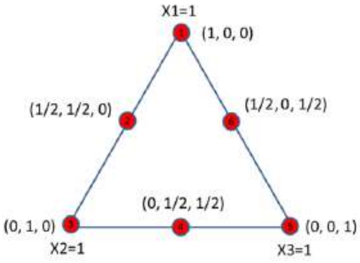
The response in a mixture experiment is usually described by a polynomial function. This function represents how the components affect the response. To get a better idea about the shape of the response surface, the natural choice for a design would be the one whose points are spread evenly over the whole simplex. An ordered arrangement consisting of a uniformly spaced distribution of points on a simplex is known as a **lattice**.

A {*q*, *m*} simplex lattice design for *q* components consists of points defined by the following coordinate settings: the proportions assumed by each component take the *m*+1 equally spaced values from 0 to 1,

$$x_i = 0, \frac{1}{m}, \frac{2}{m}, ..., 1 \;\; i = 1, 2, ..., q$$

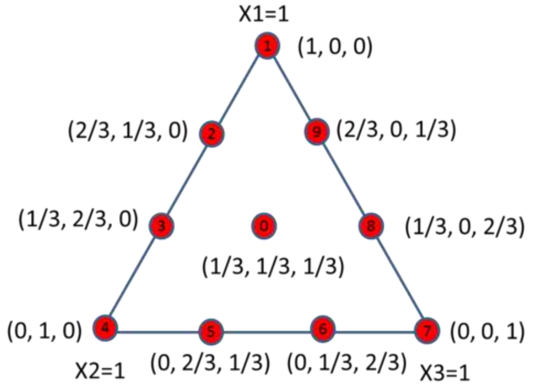
and the design space consists of all the reasonable combinations (that is summing up to 1) of values of the components. "*m*" is usually called the **degree of the lattice**. Each reasonable combination of values defines a support point of the design.

For example, for a {3, 2} design, $x_i = 0, \frac{1}{2}, 1, i = 1,2,3,$ and its design space has 6 support points. They are:



Since the {3, 2} design has 6 support points, it can be used to fit upto a quadratic response function, which also has 6 coefficients.

For a {3, 3} design, $x_i = 0, \frac{1}{3}, \frac{2}{3}, 1, i = 1, 2, 3,$, and its design space has 10 support points. They are:



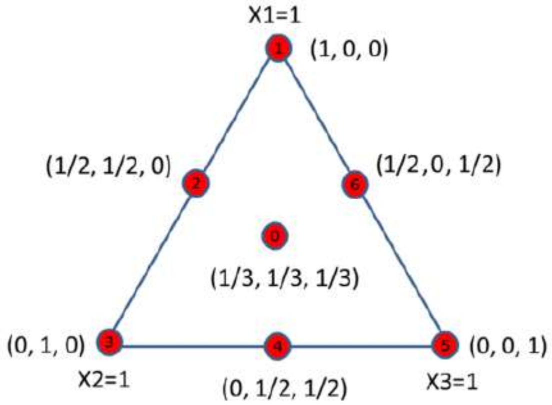
This design can be used to fit upto a full cubic response function.

In general, for a simplex design with degree *m*, where each component has *m* + 1 possible values, the experiment results can be used to fit a polynomial equation up to an order of *m*.

For a {*q*, *m*} design, the total number of support points is $\begin{pmatrix} q+m-1 \\ m \end{pmatrix}$. To reduce the number of points and still be able to fit a high order polynomial model, we often use the simplex centroid design.

Simplex Centroid Design

In a simplex centroid design, the non-zero co-ordinates of a support point have the same value. For example, the support points of a simplex centroid design for a three component mixture are as follow:



In the above simplex plot, the points (2), (4) and (6) are called the second degree centroids. Each of them has two non-zero components with equal values. Point 0 is a third degree centroid and all the three components have the same value. For a design with *q* components, the highest degree of centroid is *q*. It is called the overall centroid, or the center point of the design.

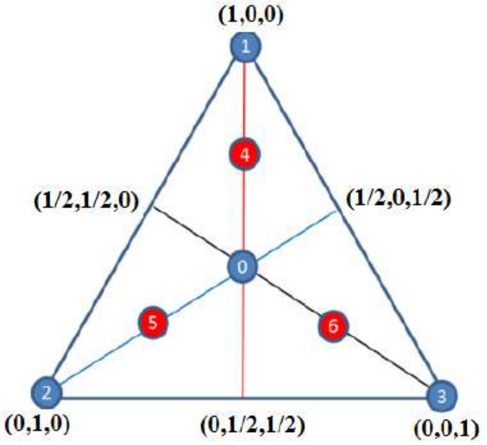
For a q component simplex centroid design, the total number of support points is $2^q - 1$. They are the points correspond to the q permutations of (1, 0, 0,…, 0), $\binom{q}{2}$ permutations of (1/2, 1/2, 0, 0, 0, 0, …,0), the $\binom{q}{3}$ permutations of (1/3, 1/3, 1/3, 0, 0, 0, 0,…, 0)…., and the overall centroid (1/q, 1/q, …, 1/q). If the degree of centroid is defined as m (< q), then the total number of support points is $\binom{q}{1}+\binom{q}{2}+...+\binom{q}{m}$.

Simplex Axial Design

The simplex lattice and simplex centroid designs have support points on the boundaries of the simplex (namely, vertices, edges, faces, etc.), except for the overall centroid. Axial designs, on the other hand, are designs consisting mainly of the points positioned inside the simplex. Axial designs have been recommended for use when the component effects are to be measured in a screening experiment, particularly when first degree models are to be fitted.

The axial of a component i is defined as the imaginary line extending from the base point $x_i = 0, x_j = 1/(q-1)$, for all $i \neq j$, to the vertex where $x_i = 1, x_j = 0$, all for all $i \neq j$.

In a simplex axial design, all the points are on the axial. The simplest form of axial design is one whose points are positioned equidistant from the overall centroid $(1/q, 1/q, ..., 1/q)$.. Traditionally, points located at half the distance from the overall centroid to the vertex are called axial points/blends. An example is given below for a three component mixture.



The points (4), (5) and (6) are the axial mixtures or blends.

A design D is specified by its support points and the replication of the support points in the experimentation. Once the design D is decided upon, the experimenter carries out the experiment say N (pre-determined) times, using the support points of the design. Suppose D has k (≥ the number of parameters in the model) support points, and the i-th point is used in n_i experiments, such that $\sum_{i=1}^{k} n_i = N$. To estimate the parameters of the response model, the method of least squares is used, based on the response observations obtained from the experimentation.

We can write the response model as Y_x = $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}+\varepsilon,$ where $\boldsymbol{\theta}$ denotes the vector of model parameters. For example, in Scheffé's linear (homogeneous) model, $\boldsymbol{f}'(\boldsymbol{x})=(x_1,x_2,...,x_q)$ and $\boldsymbol{\theta}=(\beta_1,\beta_2,...,\beta_q)'$; in Scheffé's quadratic model,

$\boldsymbol{f}'(\boldsymbol{x})=(x_1,x_2,...,x_q,x_1x_2,...,x_1x_q,...,x_{q-1}x_q)$ and
$\boldsymbol{\theta}=(\beta_1,\beta_2,...,\beta_q,\beta_1\beta_2,...,\beta_1\beta_q,...,\beta_{q-1},\beta_q)'.$

If $\boldsymbol{x}_{(i)}, i=1,2,...,k$ be the support points and $Y^{N\times 1}$ be the observed response vector, the least squares estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}=(X'_D X_D)^{-1}X'_D Y,$$

where $X_D=\left[\underbrace{\boldsymbol{f}(\boldsymbol{x}_{(1)}),...,\boldsymbol{f}(\boldsymbol{x}_{(1)})}_{n_1\,times},\underbrace{\boldsymbol{f}(\boldsymbol{x}_{(2)}),...,\boldsymbol{f}(\boldsymbol{x}_{(2)})}_{n_2\,times},...,\underbrace{\boldsymbol{f}(\boldsymbol{x}_{(k)}),...,\boldsymbol{f}(\boldsymbol{x}_{(k)})}_{n_k\,times}\right],$

and $Disp.(\hat{\boldsymbol{\theta}})=\sigma^2(X'_D X_D)^{-1}.$

To compare two designs D_1 and D_2 on a meaningful basis we consider the information matrix $X'_D X_D$ on a per observation basis, namely :

$$M_D\ = \mathbf{(}X_D{}'X_D\mathbf{)}/\ N\mathbf{.}$$

The optimum design is obtained so as to minimize some real-valued concave function of the dispersion matrix of $\hat{\boldsymbol{\theta}}$. In the class of N- point designs, it is difficult to find such a design, particularly when N is not small, as standard optimization techniques based on calculus cannot be applied. To overcome this, Kiefer introduced the concept of: ***Approximate/Continuous design***.

Continuous Design: A continuous design is characterized by its distinct support points and their masses or weights:

$$\xi=\begin{Bmatrix}\boldsymbol{x}_{(1)} & \boldsymbol{x}_{(2)} & \cdots & \boldsymbol{x}_{(k)}\\ w_1 & w_2 & \cdots & w_k\end{Bmatrix},$$

where $w_i\ \geq 0$ is the mass/weight attached to the support point $\boldsymbol{x}_{(i)}$, i = 1,2,...,k, such that $\sum\limits_{i=1}^{q} w_i=1.$

By the mass of a support point we mean the proportion of times the experiment should be conducted using that support point. For example, if the mass of a support point is 0.25 and the total number of times the experiment is repeated is 100, then the number of times the mixing proportion given by the support point is used is 100×0.25 = 4.

A commonly used continuous design, called the weighted centroid design, is defined as follows:

Weighted Centroid Design: A weighted centroid design has the same support points as a simplex centroid design, but different masses are attached to groups of points of the same type. That is, it attaches a non-negative mass w_1 to each of the support points having one non-zero co-ordinate, a non-negative mass w_2 to each of the support points having two non-zero components, and so on such that the sum of all masses is equal to 1. For example, in a two-component mixture experiment conducted using a weighted centroid design, a mass w_1 is attached to each of (1,0) and (0,1), and a mass w_2 to (1/2,1/2) such that $w_i \geq 0$ for i =1, 2, and $2w_1 + w_1 =1$. In a three-component mixture experiment, the weighted centroid design attaches a mass w_1 to (1,0,0), (0,1,0), (0,0,1), a mass w_2 to each of (1/2.1/2.0), (0,1/2,1/2), (0,1/2,1/2) and a mass w_3 to (1/3,1/3,1/3) such that $w_i \geq 0$ for i =1, 2, 3 and $3w_1 + w_2 + w_3 =1$.

To get an optimum design among the class of competing designs, two commonly used concave functions of the dispersion matrix of $\hat{\theta}$ are the determinant of $Disp.(\hat{\theta})$ and Trace [$Disp.(\hat{\theta})$]. Accordingly we get the following optimality criteria:

(a) D-optimality criterion: Minimize $| Disp.(\hat{\theta}) |$ or maximize $| X'_D X_D |$

(b) Trace-optimality criterion: Minimize Trace [$Disp.(\hat{\theta})$].

For Scheffé’s linear (homogeneous) model for a q- component mixture, the D-optimality and Trace-optimality criteria give the same optimum design, which has support points at the vertices of the simplex with equal mass. The number of support points here is equal to the number of parameters to be estimated. Such a design is called a saturated design.

For Scheffé’s quadratic model for a q- component mixture, the D-optimal design is a saturated design with support points at the vertices of the simplex and at the mid-points of the edges, each having the same mass. The trace optimal design, on the other hand is found to be a weighted centroid design with masses that minimize Trace [$Disp.(\hat{\theta})$].

MULTIVARIATE ANALYSIS OF VARIANCE

The meaning of ANOVA and MANOVA is Analysis of Variance and Multivariate Analysis of Variance, respectively. Here we shall discuss ANOVA and MANOVA in brief with their applications in agricultural science.

1. ANOVA

Analysis of Variance (ANOVA) is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned and is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the *response variable is continuous* in nature, whereas the *predictor variables are categorical*. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In a particular case, where two population means are being compared, ANOVA is equivalent to the independent two-sample *t*-test.

There are three conceptual classes of ANOVA models:

a) **Fixed-effects models**: The fixed-effects model of ANOVA applies to situations in which the experimenter applies several treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole. In it factors are fixed and are attributable to a finite set of levels of factor eg. Sex, year, variety, fertilizer etc.

   Consider for example a clinical trial where three drugs are administered on a group of men and women some of whom are married and some are unmarried. The three classifications of sex, drug and marital status that identify the source of each datum are known as factors. The individual classification of each factor is known as levels of the factors. Thus, in this example there are 3 levels of factor drug, 2 levels of factor sex and 2 levels of marital status. Here all the effects are fixed.

b) **Random effects models:** Random effects models are used when the treatments are not fixed. This occurs when the various treatments (also known as factor levels) are sampled from a larger population. When factors are random, these are generally attributable to infinite set of levels of a factor of which a random sample are deemed to occur *eg.* research stations, clinics in Delhi, sire, etc. Suppose new inject-able insulin is to be tested using 15 different clinics of Delhi state. It is reasonable to assume that these clinics are random sample from a population of clinics from Delhi.

c) **Mixed-effect models:** It describe the situations where both fixed and random effects are present.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of *t*-tests: a comparison of differences of means across more than two groups.

The ANOVA is valid under certain assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2.
- Effects are additive in nature.
- Populations have equal variance.
- Samples are randomly and dependently distributed $e_{ij} \sim$ N $(0, \sigma^2)$.

The ANOVA is performed as One-way, Two-way, three-way, etc. ANOVA when the number of factors is one, two or three respectively. In general if the number of factors is more than we perform multi-factor ANOVA.

2. Multivariate Analysis of Variance (MANOVA)

Multivariate analysis of variance (MANOVA) is a generalized form of univariate ANOVA with several dependent variables. Multivariate analysis of variance is simply an ANOVA with several dependent variables. When more than one dependent variable is studied simultaneously to see the effects of the factors (groups) then the technique of analysis used is called MANOVA. Thus MANOVA is an extension of ANOVA. Also, MANOVA is the multivariate analogue to Hotelling's T^2. The purpose of MANOVA is to test whether the vectors of means for the two or more groups are sampled from the same sampling distribution. Just as Hotelling's T^2 will provide a measure of the likelihood of picking two random vectors of means out of the same hat, MANOVA gives a measure of the overall likelihood of picking *two or more* random vectors of means out of the same hat.

For example in varietal trials the data is collected on several plant characteristics and quality parameters. In these experimental situations the data is generally analyzed separately for each of the characters. The best treatment or genotype is identified separately for each of the characters. In these situations, Multivariate Analysis of Variance (MANOVA) can be helpful. Similarly, a researcher is interested to examine the effect breed and sex in body weight, body length. Then MANOVA is applied by taking body weight and body length simultaneously as dependent variables and breed and sex as two factors.

There are two major situations in which MANOVA is used. The first is when there are several correlated dependent variables and the researcher desires a single, overall statistical test on this set of variables instead of performing multiple individual tests. The second and in some cases, the more important purpose is to explore how independent variables influence some patterning of response on the dependent variables.

- The pattern of analysis of a MANOVA is similar to ANOVA
- If there is a significant multivariate effect then examine the univariate effects (i.e. ANOVA for each dependent variable separately)
- If there is a significant univariate effect then conduct post hoc tests as necessary

Assumptions of MANOVA

- Multivariate Normality
- The sampling distributions of the dependent variables and all linear combinations of them are normal.
- Homogeneity of Variance-Covariance Matrices
- It is assumed that linear relationships between all pairs of DVs exist
- Multicollinearity – the relationship between pairs of variables is high (r>.80)
- Singularity – a variable is redundant; a variable is a combination of two or more of the other variables.

Consider a two-way MANOVA with factors as Factor A and Factor B for experiment conducted to compare v levels of Factor A and r levels of Factor B and the data is collected on p-variables. Let y_{ijk} denote the observed value of the kth response variable for the ith level of Factor A in the jth level of Factor B, $i = 1, 2, ..., v;\ j = 1, 2, ..., r; k = 1, 2, ..., p$. The data is rearranged as follows:

← Factor B →

↓ Factor A	1	2	…	j	…	r	**Factor A Mean ↓**
1	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	…	$\mathbf{y}_{1j}$	…	$\mathbf{y}_{1r}$	$\bar{\mathbf{y}}_{1.}$
2	$\mathbf{y}_{21}$	$\mathbf{y}_{22}$	…	$\mathbf{y}_{2j}$	…	$\mathbf{y}_{2r}$	$\bar{\mathbf{y}}_{2.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
i	$\mathbf{y}_{i1}$	$\mathbf{y}_{i2}$	…	$\mathbf{y}_{ij}$	…	$\mathbf{y}_{ir}$	$\bar{\mathbf{y}}_{i.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
v	$\mathbf{y}_{v1}$	$\mathbf{y}_{v1}$	…	$\mathbf{y}_{v1}$	…	$\mathbf{y}_{v1}$	$\bar{\mathbf{y}}_{v.}$
Factor B Mean→	$\bar{\mathbf{y}}_{.1}$	$\bar{\mathbf{y}}_{.2}$	…	$\bar{\mathbf{y}}_{.j}$	…	$\bar{\mathbf{y}}_{.r}$	$\bar{\mathbf{y}}_{..}$

Here $\mathbf{y}_{ij} = (\,y_{ij1}\ \ y_{ij2}\ ...y_{ijk}\ ...\ y_{ijp})$ is a p-variate vector of observations.

$$\bar{\mathbf{y}}_{i.} = \frac{1}{r}\sum_{j=1}^{r}\mathbf{y}_{ij}\ ;\ \bar{\mathbf{y}}_{.j} = \frac{1}{v}\sum_{i=1}^{v}\mathbf{y}_{ij}\ \text{ and }\ \bar{\mathbf{y}}_{..} = \frac{1}{vr}\sum_{i=1}^{v}\sum_{j=1}^{r}\mathbf{y}_{ij}\ .$$

The observations can be represented by a two-way classified multivariate model $\mathbf{\Omega}$

$$\mathbf{\Omega}:\ \mathbf{y}_{ij} = \boldsymbol{\mu} + \mathbf{t}_i + \mathbf{b}_j + \mathbf{e}_{ij} \qquad i = 1, 2,\ldots,v;\ j = 1, 2,\ldots,b, \qquad (1)$$

$\boldsymbol{\mu}$ = $(\mu_1\ \mu_2 \ldots \mu_k \ldots \mu_p)'$ is the $p \times 1$ vector of general means, $t_i = (t_{i1}\ t_{i2} \ldots t_{ik} \ldots t_{ip})'$ are the effects of ith level of Factor A on p-characters, and $b_j=(b_{j1}\ b_{j2} \ldots b_{jk} \ldots b_{jp})'$ are the effects of jth level of Factor B on p-characters. $\mathbf{e}_{ij} = (e_{ij1}\ e_{ij2} \ldots e_{ijk} \ldots e_{ijp})'$ is a p-variate random vector associated with $\mathbf{y}_{ij}$ and assumed to be distributed independently as p variate normal distribution $\mathbf{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The equality of treatment effects is to be tested i.e. H_0: $(t_{i1}\ t_{i2}\ldots t_{ik}\ldots t_{ip})'$ = $(t_1\ t_2\ldots t_k\ldots t_p)'$ (say) $\forall i = 1,2,\cdots,p$ against the alternative $H_1:$ at least two of the Factor A effects are unequal. Under the null hypothesis, the model (1) reduces to

$$\mathbf{\Omega}_0 : \mathbf{y}_{ij} = \boldsymbol{\alpha} + \mathbf{b}_j + \mathbf{e}_{ij} \qquad (2)$$

where $\boldsymbol{\alpha} = (\mu_1 + t_1\ \mu_2 + t_2, ..., \mu_p + t_p)'$.

An outline of MANOVA Table for testing the equality of treatment effects and replication effects is

MANOVA Table

Source	**DF**	**SSCPM (Sum of Squares and Cross Product Matrix)**
Factor A	v-1 = h	**H =** $b\sum_{i=1}^{v}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})'$
Factor B	r-1 = t	**B=** $v\sum_{j=1}^{b}(\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{.j} - \bar{\mathbf{y}}_{..})'$
Residual	(v-1)(r-1) = s	**R=** $\sum_{i=1}^{v}\sum_{j=1}^{b}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{.j} + \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{.j} + \bar{\mathbf{y}}_{..})'$
Total	vr-1	**T=** $\sum_{i=1}^{v}\sum_{j=1}^{b}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})'$ **=H+B+R**

Here **H**, **B**, **R** and **T** are the sum of squares and sum of cross product matrices of Factor A, Factor B, errors (residuals) and totals respectively. The residual sum of squares and cross products matrix for the reduced model Ω_0 is denoted by $\mathbf{R}_0$ and is given by $\mathbf{R}_0 = \mathbf{R} + \mathbf{H}$.

The null hypothesis of equality of treatment mean vectors is rejected if the ratio of generalized variance (*Wilk's lambda* statistic) $\Lambda = \dfrac{|\mathbf{R}|}{|\mathbf{H}+\mathbf{R}|}$ is too small. Assuming the normal distribution, Rao (1973) showed that under null hypothesis Λ is distributed as the product of independent beta variables. A better but more complicated approximation of the distribution of Λ is

$$\frac{1-\Lambda^{1/b}}{\Lambda^{1/b}}\frac{(ab-c)}{ph} \sim \text{F (ph, ab-c)}$$

where $a = \left(s - \dfrac{p-h+1}{2}\right)$, $b = \sqrt{\left\{\left(p^2h^2-4\right)/\left(p^2+h^2-5\right)\right\}}$, $c = \dfrac{ph-2}{2}$

For some particular values of h and p, it reduces to exact F-distribution. The special cases are given below:

For h = 1 and any *p*, this reduces to $\dfrac{(1-\Lambda)}{\Lambda}\dfrac{(s-p+1)}{p} \sim$ F (p, s – p + 1)

For h=2 and any p, it reduces to $\dfrac{(1-\sqrt{\Lambda})}{\sqrt{\Lambda}}\dfrac{(s-p+1)}{p} \sim$ F (2p, 2(s – p + 1))

For p=2 and any h: $\frac{(1-\sqrt{\Lambda})}{\sqrt{\Lambda}}\frac{(s-1)}{h}$ ~ F (2h, 2(s – 1)).

For p = 1, the statistic reduces to the usual variance ratio statistics.

The hypothesis regarding the equality of Factor B effects can be tested by replacing Λ by $\frac{|\mathbf{R}|}{|\mathbf{B}+\mathbf{R}|}$ and h by t in the above.

Several other criteria viz. Pillai's Trace, Hotelling-Lawley Trace or Roy's Greatest Root are available in literature for testing the null hypothesis in MANOVA. Wilks' Lamda is, however, the commonly used criterion. Here, we shall restrict to the use of Wilks' Lamda criterion. For further details on MANOVA, a reference may be made to Seber (1983) and Johnson and Wichern (1988).

Remark 1: One complication of multivariate analysis that does not arise in the univariate case is the ranks of the matrices. The rank of **R** should not be smaller than p or in other words error degrees of freedom s should be greater than or equal to p (s ≥ p).

Advantages of MANOVA

In comparison to ANOVA, MANOVA has the following advantages:

- The researcher improves their chances of finding what changes as a result of the experimental treatment
- Since only ‘one’ DV is tested the researcher is protected against inflating the type 1 error due to multiple comparisons
- It can show differences that individual ANOVAs do not – it is sometimes more powerful

2.1 Multivariate Treatment Contrast Analysis

If the treatments are found to be significantly different through MANOVA, then the next question is “which treatments are significantly different?” This question can be answered through multivariate treatment contrast analysis. In the literature, the multivariate treatment contrast analysis is generally carried out using the χ^2-statistic. The χ^2-statistic is based on the assumption that the error variance-covariance matrix is known. The error variance-covariance matrix is, however, generally unknown. Therefore, the estimated value of error variance-covariance matrix is used. The error variance-covariance matrix is estimated by sum of squares and cross products (SSCP) matrix for error divided by the error degrees of freedom. As a consequence, test based on χ^2-statistic is an approximate solution. The procedure using the Wilk’s Lambda criterion is also described in the sequel.

Suppose the hypothesis to be tested is H$_0$: $\mathbf{t}_i = \mathbf{t}_{i'}$ against H$_1$:$\mathbf{t}_i \neq \mathbf{t}_{i'}$. This hypothesis can be rewritten as

$$H_0: = (\mathbf{t}_i - \mathbf{t}_{i'}) = \mathbf{0} \text{ against } H_1: = (\mathbf{t}_i - \mathbf{t}_{i'}) \neq \mathbf{0}, \qquad (3)$$

where $(\mathbf{t}_i - \mathbf{t}_{i'})' = \left(t_{i1}-t_{i'1} \quad t_{i2}-t_{i'2} \quad \ldots \quad t_{ik}-t_{i'k} \quad \ldots \quad t_{ip}-t_{i'p}\right)$. Here t_{ik} denote the effect of treatment i for the dependent variable k. The best linear unbiased estimate of $(\mathbf{t}_i - \mathbf{t}_{i'})$ is

$$\left(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{i'.}\right)' = \left(\bar{y}_{i1}-\bar{y}_{i'1} \quad \bar{y}_{i2}-\bar{y}_{i'2} \quad \ldots \quad \bar{y}_{ik}-\bar{y}_{i'k} \quad \ldots \quad \bar{y}_{ip}-\bar{y}_{i'p}\right)$$

where $\bar{y}_{ik}$ is the mean of treatment i for variable *k*.

i) χ^2 − Test

The statistic based on χ^2, requires covariance matrix of the contrast of interest. The covariance matrix, in case of a RCB design for elementary treatment contrast is obtained by dividing the SSCP matrix for errors obtained in MANOVA by half of the product of error degrees of freedom and the number of replications. Let this variance-covariance matrix is denoted by $\boldsymbol{\Sigma}_c$. Under null hypothesis, $\mathbf{x} = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{i'.}$ follows p- variate normal distribution with mean vector **0** and variance-covariance matrix $\boldsymbol{\Sigma}_c$. Applying the Aitken's transformation, it can be shown that $\mathbf{z} = \boldsymbol{\Sigma}_c^{-1/2}\mathbf{x}$ follows a p-variate normal distribution with mean vector **0** and variance-covariance matrix $\mathbf{I}_g$, where $\mathbf{I}_g$, denotes the identity matrix of order g. Then using the results of quadratic forms, it can easily be seen that $\mathbf{z'z} = \mathbf{x'}\boldsymbol{\Sigma}^{-1}\mathbf{x}$ follows a χ^2 distribution with p-degrees of freedom.

ii) Wilk's Lambda Criterion

For testing the null hypothesis (3), we obtain a sum of squares and products matrix for the above elementary treatment contrast. Let the SSCP matrix for above elementary treatment contrast be $\mathbf{G}_{p\times p}$. The diagonal elements of **G** are then obtained by

$$g_{kk} = \left(\frac{r}{2}\right)(\bar{y}_{ik} - \bar{y}_{i'k})^2 \quad \forall \quad k = 1,2,...,p;\ i \neq i' = 1,2,...,v \tag{4}$$

and the off diagonal elements are obtained by

$$g_{kk'} = \frac{r}{2}(\bar{y}_{ik} - \bar{y}_{i'k})(\bar{y}_{ik'} - \bar{y}_{i'k'}) \tag{5}$$

The null hypothesis is rejected if the value of Wilk's Lambda $\boldsymbol{\Lambda}^* = \dfrac{|\mathbf{R}|}{|\mathbf{G}+\mathbf{R}|}$ is small, where **R** is the SSCP matrix due to residuals as obtained through MANOVA. The hypothesis is then tested using the following F-test statistics based on Wilk's Lambda for *h* = 1

$$\frac{1-\boldsymbol{\Lambda}^*}{\boldsymbol{\Lambda}^*}\frac{\text{edf} - p + 1}{p} \sim \text{F(p, s-p+1)}.$$

Exercise 1: An experiment was conducted at IGFRI, Jhansi to investigate the effect of four types of trees (treatments) on different parameters viz. height, collar diameter, DBH, crown diameter. The data are as follows:

TREE-TYPE	HEIGHT	COLLAR DIAMETER	DBH	CROWNDIA
1	4	10.5	6.9	15.13
1	3.6	9.3	6.3	10.21
1	1.5	2.5	1.2	0.17
1	3.8	7	4.3	2.63
2	3	9.1	5.8	5.24
2	3.7	8.1	5.1	5.89
2	3.8	7.5	5.5	4.47
2	3.8	9.8	6.4	7.57
3	5.3	11.1	6.9	13.09
3	4.9	12.1	8	12.93
3	5.6	13.7	9.2	15.26
3	4.5	10.3	6.5	10.55

4	4.7	13.7	9.1	20.66
4	4.8	14.9	10	25.62
4	4.6	11.7	9.7	16.21
4	5.5	12.7	8.7	17.79

Analyze the data given to examine the effect of tree-type on the four measurements of tree by using MANOVA and interpret the results.

CLUSTER ANALYSIS

1. Introduction

Cluster analysis is usually done in an attempt to combine cases into groups when the group membership is not known prior to the analysis. Cluster analysis is a technique for grouping individual or objects into unknown groups. It differs from other methods of classification such as Discriminant analysis, in that in cluster analysis the number and characteristics of the groups are to be derived from the data and are not usually known prior to the analysis.

In biology, cluster analysis has been used for decades in the area of taxonomy, where living things are classified into arbitrary groups on the basis of their characteristics group. The classification proceeds from the most general to the most specific in steps. The most general classification is kingdom followed by phylum, subphylum, and class etc. Cluster analysis has been used in medicine to assign patient to specific diagnostic categories on the basis of their presenting symptoms and signs. Cluster analysis is also an important tool for investigation in data mining. For example consumers can be clustered on the basis of their purchases in marketing research. Here the emphasis may be on the methods that can be used for large data sets. In short it is possible to find application of cluster analysis in virtually any field of research. It is also possible to cluster the variables rather than the cases. Clustering of variables is sometimes used in analyzing the items in a scale to determine which items tends to be close together in terms of individual response to them.

2. Clustering Methods (Johnson and Wichern, 2006)

The commonly used methods of clustering fall into two general categories.

(i) Hierarchical and
(ii) Non hierarchical.

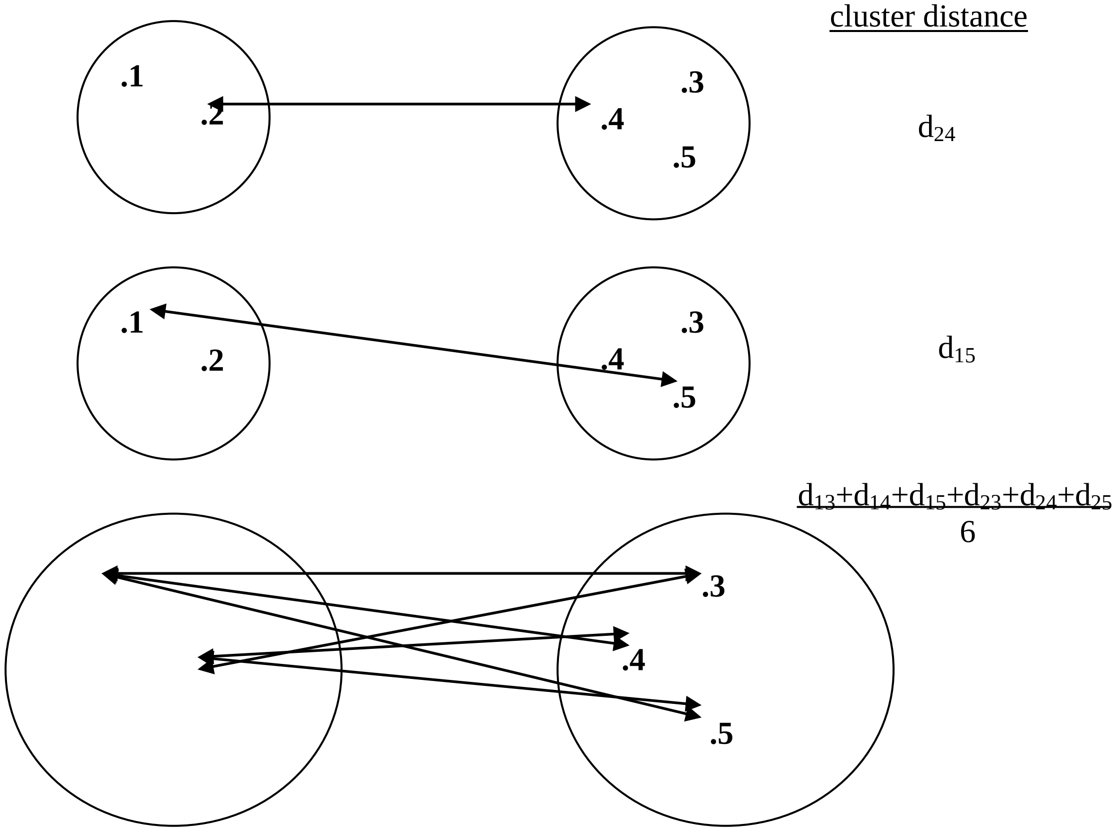
Hierarchical clustering techniques proceed by either a series of mergers or a series of successive divisions. Agglomerative hierarchical method starts with the individual objects, thus there are as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two sub groups such that the objects in one sub group are far from the objects in the others. These subgroups are then further divided into dissimilar subgroups. The process continues until there are as many subgroups as objects i.e., until each object form a group. The results of both agglomerative and divisive method may be displayed in the form of a two dimensional diagram known as Dendrogram. It can be seen that the Dendrogram illustrate the mergers or divisions that have been made at successive levels.

Linkage methods are suitable for clustering items, as well as variables. This is not true for all hierarchical agglomerative procedure. The following types of linkage are now discussed:

(i) Single linkage (minimum distance or nearest neighbour),
(ii) Complete linkage (maximum distance or farthest neighbour) and
(iii) Average linkage (average distances).

The merging of cluster under the three linkage criteria is illustrated schematically in the figure given below.



From the above figure, we see that Single linkage results when groups are fused according to the distance between their nearest members. Complete linkage occurs when groups are fused according to the distance between there farthest members. For Average linkage, groups are fused according to the average distance between pair of members in the respective sets.

The following are the steps in the agglomerative hierarchical clustering algorithm for groups of N objects (items or variables).

i. Start with N clusters, each containing a single entity and an N×N symmetric matrix of distance (or similarities) **D** = $\{d_{ik}\}$.
ii. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between most similar clusters U and V be $d_{uv.}$
iii. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters.
iv. Repeat steps (ii) and (iii) a total of N-1 times (All objects will be in a single cluster after the algorithm terminates). Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

The basic ideas behind the cluster analysis are now shown by presenting the algorithm components of linkage methods.

2.1 Single Linkage

The inputs to a single linkage algorithm can be distances or similarities between pair of objects. Groups are formed from the individual entities by merging nearest neighbors, i.e. smallest distance or largest similarities.

Initially, we must find the smallest distance in **D** = $\{d_{ik}\}$ and merge the corresponding objects, say, U and V, to get cluster (UV). For step 3 of general algorithm the distance between (UV) and any other cluster W are computed by $d_{(u,v),w} = \min \{d_{uw}, d_{vw}\}$

The results of single linkage clustering can be graphically displayed in the form of Dendrogram or tree diagram. The branches in the tree represent clusters. The branches come together (merge) at nodes whose positions along a distance (or similarity) axis indicate the level at which the fusion occurs.

2.2 Complete Linkage

Here at each stage, the distance (similarity) between clusters is determined by the distance (similarity) between the two elements. One from each cluster that is most distant. Thus complete linkage ensures that all items in a cluster are with in some maximum distance (or minimum similarity) of each other.

The general agglomerative algorithm again starts by finding the minimum entry in D = $\{d_{ik}\}$ and merging the corresponding objects, such as U and V, to get cluster (UV). For step (iii) of general algorithm, the distance between (UV) and any other cluster W is

$$D_{(uv)w} = \max \{d_{uw} , d_{vw}\}$$

Here d_{uw} and d_{vw} are the distances between the most distant members of clusters U and W and clusters V and W.

2.3 Average Linkage

Average linkage treats the distances between two clusters as the average distance between all pairs of items where one member of pair belongs to each cluster.

Again the input to average linkage algorithm may be distances or similarities and the method can be used to group objects or variables. The average linkage algorithm proceeds in the manner of the general algorithm, we begin by searching the distance matrix D = $\{d_{ik}\}$ to find the nearest (most similar) objects for example U and V. These objects are merged to form the cluster (UV). For step 3 of general agglomerative algorithm the distance between (UV) and other cluster W are determined by

$$d_{(uv)w} = (\sum_i \sum_k d_{ik}) / (N_{(uv)} * N_w),$$

where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W, and N_{uv} and N_w are the member of items in clusters (UV) and W respectively.

2.4 Centroid

This method assigns each item to the cluster having nearest centroid (means). The process has three steps,

i. Partition the items into k initial clusters.

ii. Proceed through the list of items assigning an item to the cluster whose centroid (mean) is nearest. Recalculate the centroid (mean) for the cluster receiving the new item and the cluster losing the item.
iii. Repeat step (ii) until no more assignments take place.

2.5 Ward's Hierarchical Clustering Methods

Ward considered hierarchical clustering procedure based on minimizing the loss of information from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion, ESS. First for a given cluster k, let ESS_k be the sum of the square deviation of every item of the cluster from the cluster mean (centroid). If there are currently K clusters, define ESS as the sum of the ESS_k or ESS = ESS_1+ESS_2+ …. +ESS_k. At each step in the analysis the union of every possible pair of cluster is considered and the two clusters whose combination results in the smallest increase in ESS (minimum loss of information) are joined. Initially each cluster consist of a single item, and if there are N items, ESS_k = 0, k = 1, 2,…, N so ESS = 0 at the other extreme, when all the clusters are combined in a single group of N items, the value of ESS is

$$\text{ESS} = \sum_{j=1}^{N}(X_j - \overline{X})'(X_j - \overline{X}),$$

where X_j is the multivariate measurement associated with the j^{th} item and $\overline{X}$ is the mean of all the items. The results of Ward's method can be displayed by a Dendrogram. The vertical axis gives the value of ESS at which the mergers occur.

2.6 Non Hierarchical Clustering Method

Non hierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure. Because a matrix of distance does not have to be determined and the basic data do not have to be stored during the computer run. Non hierarchical methods can be applied to much larger data sets than can hierarchical techniques. Non hierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points which will form nuclei of the cluster.

2.7 K means Clustering (Afifi, Clark and Marg, 2004)

The K means clustering is a popular non hierarchical clustering technique. For a specified number of clusters K the basic algorithm proceeds in the following steps.

i. Divide the data into K initial cluster. The number of these clusters may be specified by the user or may be selected by the program according to an arbitrary procedure.

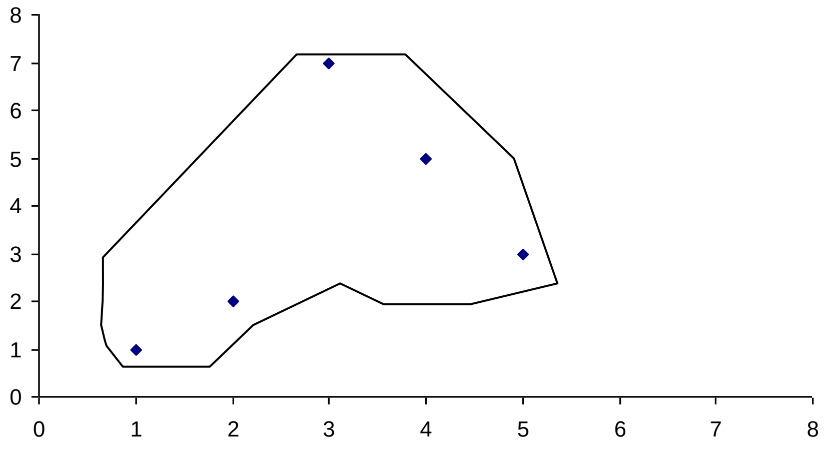
ii. Calculate the means or centroid of the K clusters.

iii. For a given case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.

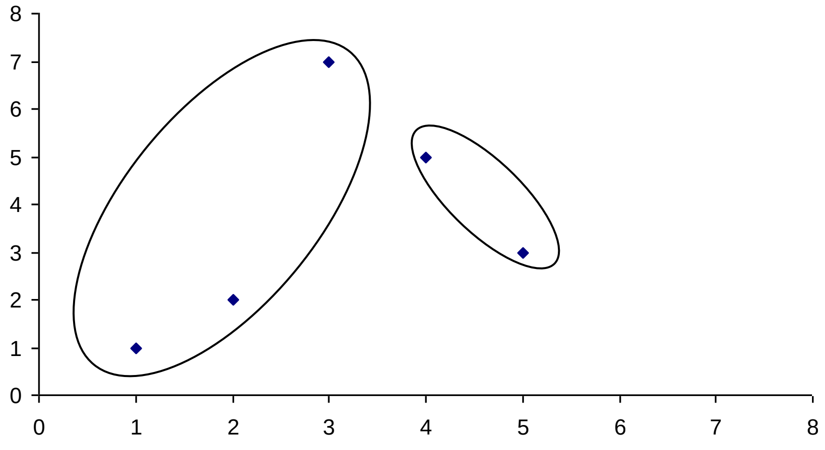
iv. Repeat step (iii) for each case.

v. Repeat steps (ii), (iii), and (iv) until no cases are reassigned.

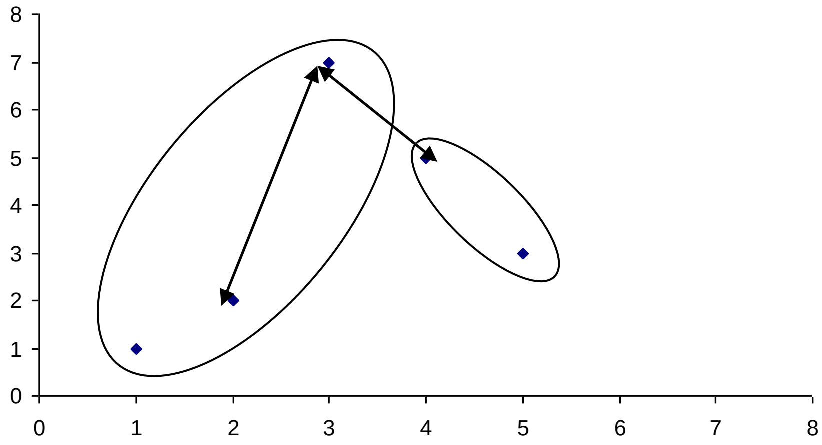
The first step considers all the data as one cluster. For the hypothetical data set this step is illustrated as in the figure below. The algorithm then searches for the variable, with the highest variance in this case X_1. The original cluster is now split into two clusters using the mid range of X_1 as the dividing point as shown in plot (b) of figure drawn below. If the data are standardized, then each variable has a variance of one. In that case the variable with the smallest range is selected to make the split. The algorithm in general proceeds in this manner by further splitting the clusters until the specified member K is achieved. That is, it successively finds that particular variable and the cluster producing the largest variance and splits that cluster accordingly until K clusters are obtained. At this stage, step (i) of the basic algorithm is completed and it proceeds with the other steps.



(a) Starts with all points in one cluster.



(b) Cluster is split into 2 clusters at mid range of X_1(variable with largest var.)



(c) Point 3 is closure to centroid of cluster (1, 2, 3) and stays assigned to (1, 2, 3)

(d) Every point is now closest to centroid of its own cluster.

3. Dendrogram

Dendrogram is also called hierarchical tree diagram or plot, and shows the relative size of the proximity coefficients at which cases are combined. The bigger the distance coefficient or the smaller the similarity coefficient, the more clustering involved combining unlike entities, which may be undesirable. Trees are usually depicted horizontally, not vertically, with each row representing a case on the Y axis, while the X axis is a rescaled version of the proximity coefficients. Cases with low distance/high similarity are close together. Cases showing low distance are close, with a line linking them a short distance from the left of the Dendrogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness. When, on the other hand, the linking line is to the right of the Dendrogram the linkage occurs at a high distance coefficient, indicating the cases/clusters were agglomerated even though much less alike. If a similarity measure is used rather than a distance measure, the rescaling of the X axis still produces a diagram with linkages involving high alikeness to the left and low alikeness to the right.

4. Proximity Measures (Timm, 2002)

Proximity measures are used to represent the nearest of two objects. If a proximity measure represents similarity, the value of the measure increases as two objects become more similar. Alternatively if the proximity measure represents dissimilarities the value of the measure decreases in value as two objects become more alike. Let X and Y represents two objects in a p-variate space then an example of dissimilarity measures is the Euclidian distance between X and Y. For measure of similarity, we may use the proportion of the elements in the two vectors that match.

4.1 Dissimilarity Measures

Given two objects X and Y in a ‘p’ dimensional space, a dissimilarity measure satisfies the following conditions:

1. $d\,(X,Y) \geq 0$ for all objects X and Y.
2. $d\,(X,Y) = 0$ iff $X = Y$.
3. $d\,(X,Y) = d\,(Y,X)$.

Condition (3) implies that the measure is symmetric so that the dissimilarity measure that compares X and Y is same as the comparison for object Y verses X. Condition (2) requires the measures to be zero, when ever object X equals to object Y. The objects are identical if d(X, Y) = 0. Finally, Condition (1) implies that the measure is never negative.
Some dissimilarity measures are as follows.

4.1.1 Euclidian Distance

This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as,

$$d(X,Y)= \{\sum_{i=1}^{p}(X_i - Y_i)^2\}^{1/2} \quad \text{or}$$

in matrix form

$$d\,(X,Y)= \sqrt{(X-Y)'(X-Y)}$$

where $\mathbf{X}' = (X_1, X_2, \ldots , X_p)$
$\mathbf{Y}' = (Y_1, Y_2, \ldots , Y_p)$

The statistical distance between the same two observations is of the form

$$d\,(X,Y) = \sqrt{(X-Y)'A(X-Y)}\,,$$

where $\mathbf{A} = \mathbf{S}^{-1}$ and S contains the sample variances and covariances.

Euclidian and square Euclidian distances are usually computed from raw data and not from standardized data.

4.1.2 Square Euclidean Distance

Square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as:

$$d^2(X,Y) = \sum_{i=1}^{p}(X_i - Y_i)^2$$

or in matrix form

$$d^2(X,Y) = (\mathbf{X} - \mathbf{Y})' (\mathbf{X} - \mathbf{Y})$$

4.1.3 Minkowski Metric

When there is no idea about prior knowledge of the distance group then one goes for minkowski metric. This can be computed as given below:

$$d(X,Y) = \{\sum_{i=1}^{p}|X_i - Y_i|^m\}^{\frac{1}{m}}$$

For m = 1, d(X,Y) measures the city block distance between two points in p dimensions. For m = 2, d(X,Y) becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

4.1.4 City-Block (Manhattan) Distance

This distance is simply the average difference across dimensions. In most cases, this distance measure yields result similar to the simple Euclidean distance. This can be computed as :

$$d(X,Y) = \sum_{i=1}^{p} \left| X_i - Y_i \right|$$

4.1.5 Chebychev Distance

This distance measure may be appropriate in case when we want to define the objects as different if they are different on any one of the dimensions. The chebychev distance is computed as:

$$d(X,Y) = \text{maximum} \left| X_i - Y_i \right|$$

Two additional popular measures of distance or dissimilarity are given by the Canberra metric and the Czekanowski coefficient. Both of these measures are defined for non negative variables only. We have

Canberra Metric: $d(X, Y) = \sum_{i=1}^{p} \frac{\left| X_i - Y_i \right|}{(X_i + Y_i)}$

Czekanowski Coefficient = $1 - \frac{2\sum_{i=1}^{p} \min(X_i, Y_i)}{\sum_{i=1}^{p} (X_i - Y_i)}$

4.2 Similarity Measure

Given two objects X and Y in a p-dimensional space, a similarity measure satisfies the following conditions:

1. $0 \leq S(X,Y) \leq 1$ for all objects X and Y
2. S(X,Y) = 1 iff X = Y
3. S(X,Y) = S(Y, X)

Here S(X,Y) = 1 – d(X,Y)
S(X,Y) = similarity measure
D(X,Y) = dissimilarity measure

Let the frequency of matches and mix matches for objects X and Y be arranged in the form of a contigency table as follows:

		Object (X)		
		1	0	Totals
Object(Y)	1	a	b	a+b
	0	c	d	c+d
Totals		a+c	b+d	P=a+b+c+d

a represents the frequency of 1-1 matches
b represents the frequency of 1-0 matches

c represents the frequency of 0-1 matches
d represents the frequency of 0-0 matches

Following is the list of common similarity coefficients defined in terms of the frequency in the table.

	Coefficient	Rationale
1.	$(a+d)/p$	Equal weights for 1-1 matches and 0-0 matches.
2.	$2(a+d)/(2(a+d)+b+c)$	Double weight for 1-1 matches and 0-0 matches.
3.	$(a+d)/(a+d+2(b+c))$	Double weight for unmatched pairs.
4.	a/p	No 0-0 matches in numerator.
5.	$a/(a+b+c)$	No 0-0 matches in numerator or denominator.
6.	$2a/(2a+b+c)$	No 0-0 matches in numerator and denominator. Double weight for 1-1 matches
7.	$a/(a+2(b+c))$	No 0-0 matches in numerator or denominator. Double weight for unmatched pairs
8.	$a/(b+c)$	Ratio of matches to mismatches with 0-0 Matches excluded.

Coefficient of 1, 2, and 3 in the table are monotonically related. Suppose coefficient-1 is calculated for two contingency table. If $[(a_1+d_1)/p] \geq [(a_{11}+d_{11})/p]$, then we also have $[2(a_1+d_1)/(2(a_1+d_1)+b_1+c_1)] \geq [2(a_{11}+d_{11})/(2(a_{11}+d_{11})+b_{11}+c_{11})]$ and coefficient 3 will be at least as large for table 1 as it is for table 2.

Here a_1, b_1, c_1, d_1 are from table 1 and a_{11}, b_{11} ,c_{11} , d_{11} are from table 2.

5. Illustration (Chatfield and Collins, 1990)

Given below is food nutrient data on calories, protein, fat, calcium and iron. The objective of the study is to identify suitable clusters of food nutrient data based on the five variables**.**

Food Items	Calories	Protein	Fat	Calcium	Iron
1	340	20	28	9	2.6
2	245	21	17	9	2.7
3	420	15	39	7	2
4	375	19	32	9	2.6
5	180	22	10	17	3.7
6	115	20	3	8	1.4
7	170	25	7	12	1.5
8	160	26	5	14	5.9

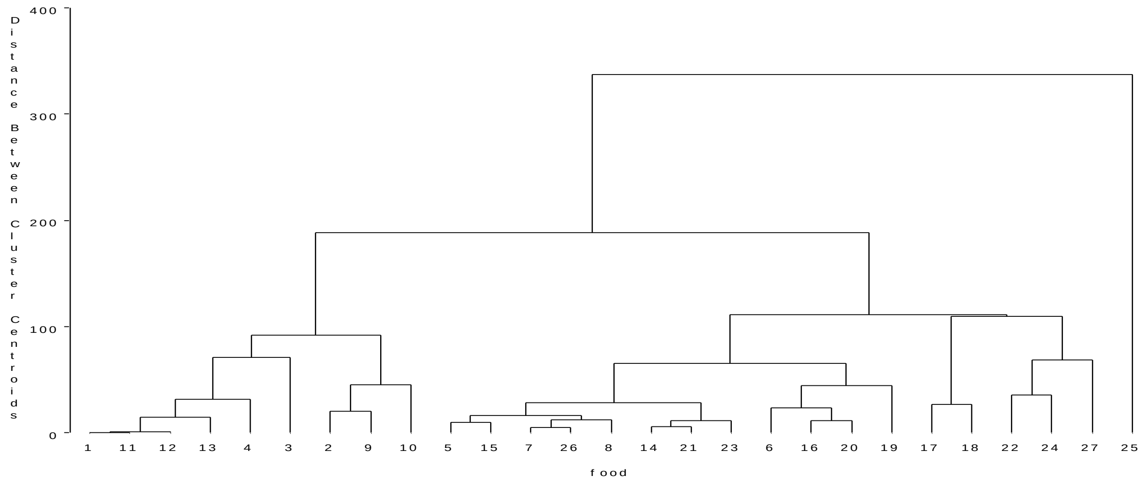
9	265	20	20	9	2.6
10	300	18	25	9	2.3
11	340	20	28	9	2.5
12	340	19	29	9	2.5
13	355	19	30	9	2.4
14	205	18	14	7	2.5
15	185	23	9	9	2.7
16	135	22	4	25	0.6
17	70	11	1	82	6
18	45	7	1	74	5.4
19	90	14	2	38	0.8
20	135	16	5	15	0.5
21	200	19	13	5	1
22	155	16	9	157	1.8
23	195	16	11	14	1.3
24	120	17	5	159	0.7
25	180	22	9	367	2.5
26	170	25	7	7	1.2
27	170	23	1	98	2.6

R-Code for Performing Cluster Analysis Based on the Above data

Following R code is useful for the above problem. Here, k=3 has been mentioned for getting three clusters. For getting more clusters, accordingly number need to be changed.

```
rw<-read.csv(file.choose(),header = TRUE) #data entry from CSV
rw
rw1<-as.matrix(rw)
rw1
row.names(rw1)<-seq(1:27) # name of the row for which grouping need to be done
rw1
rw2<-as.data.frame(scale(rw1)) #for standarization
rw2
install.packages(c("cluster", "factoextra")) #required package
library (cluster)
library(factoextra)
summary(rw2)
dist_mat<-dist(rw2, method = 'euclidian') #for distance matrix
dist_mat
#dendogram
hclust_avg <- hclust(dist_mat, method = 'average')
plot(hclust_avg)
plot(hclust_avg)
rect.hclust(hclust_avg , k = 3, border = 2:6)
abline(h = 2, col = 'red')
```

Dendrogram for above data



Interpretation

The main objective of our analysis is to grouping the food items on the basis of their nutrient content based on the five variables such that food items with in the groups are homogeneous and between the groups are heterogeneous.

Number of groups	Food items
Two groups	Group-1 (1,11,12,…,27) Group-2 (25)
Three groups	Group-1 (1,11,…,10) Group-2 (5,15,…,27) Group-3 (27)
Four groups	Group-1 (1,11,…,10) Group-2 (5,15,…,19) Group-3 (17,18,…,27) Group-4 (25)
Five groups Five groups	Group-1 (1,11,…,10) Group-2 (5,15,…,19) Group-3 (17,18) Group-4 (22,24,27) Group-5 (25)
Six groups	Group-1 (1,11,…,3) Group-2 (2,9,10) Group-3 (5,15,…,19) Group-4 (17,18) Group-5 (22,24,27) Group-6 (25)

6. Examples of Clustering Application

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- **Land Use**: Identification of areas of similar land use in earth observation database.
- **Insurance**: Identifies groups motor insurance policy holders with a high average claim cost.
- **City Planning**: Identification of group of houses according to their house type, value and geographical location.

- **Earthquake Studies**: Observed earthquake epicenters should be clustered along continent faults.
- **Field of medicine**: Clustering of diseases, cure for disease of symptoms of disease can lead to very useful taxonomies.
- **Field of psychiatry**: The correct diagnosis of clusters of symptoms such as Paranoia, Schizophrenia etc. is essential for successful therapy.
- **In Archeology**: Researches have attempt to establish taxonomies of stone tools, funerals object etc by applying cluster analytic techniques.
- **Field of plant and animal ecology**: Clustering is used to describe and to make spatial and temporal comparison of communities of organism in heterogeneous environment.
- **Field of Bioinformatics:** In transcriptomics, clustering is used to build groups of genes with related represents patterns and also in sequence analysis, it is used to group homologous sequence into gene families.
- **Social network analysis**: In the study of social network, clustering may be used to recognize community with large group of people. In general, when ever we need to classify a mountain of information into manageable meaningful piles, cluster analysis is of great utility. It is also used in data mining.

7. Conclusions

In this presentation, different issues related to cluster analysis have been discussed. Unlike other methods of classification, cluster analysis however, has not yet gained a standard methodology. Nonetheless a number of techniques are developed for dividing multivariate sample on a composition which is not known in advance into several groups.

Cluster analysis is a heuristic technique for classifying cases into groups when knowledge of the actual group membership is unknown. There are numerous method for performing the analysis, with out good guidelines for choosing among them. Unless there is considerable separation among the inherent group, it is not realistic to expect very clear results with cluster analysis. In particular if the observations are distributed in a nonlinear manner, it may be difficult to achieve distinct groups. Cluster analysis is quite sensitive to outliers. In fact it is sometimes used to find outlier. The data should be carefully screened before running cluster programs. Many statistical package programs are also being used for the purpose of cluster analysis.

DISCRIMINANT FUNCTION ANALYSIS

Introduction

Discriminant function analysis is a statistical analysis to predict a categorical dependent variable (called a grouping variable) by one or more continuous or binary independent variables (called predictor variables). The original dichotomous discriminant analysis was developed by Sir Ronald Fisher in 1936. It is different from an ANOVA or MANOVA, which is used to predict one (ANOVA) or multiple (MANOVA) continuous dependent variables by one or more independent categorical variables. Discriminant function analysis is useful in determining whether a set of variables is effective in predicting category membership. Discriminant analysis is used when groups are known a priori (unlike in cluster analysis). Each case must have a score on one or more quantitative predictor measures, and a score on a group measure. In simple terms, discriminant function analysis is classification - the act of distributing things into groups, classes or categories of the same type.

Moreover, it is a useful follow-up procedure to a MANOVA instead of doing a series of one-way ANOVAs, for ascertaining how the groups differ on the composite of dependent variables. In this case, a significant F test allows classification based on a linear combination of predictor variables. Terminology can get confusing here, as in MANOVA, the dependent variables are the predictor variables, and the independent variables are the grouping variables.

Assumptions

The assumptions of discriminant analysis are the same as those for MANOVA. The analysis is quite sensitive to outliers and the size of the smallest group must be larger than the number of predictor variables. The major assumptions are:

- Multivariate normality: Independent variables are normal for each level of the grouping variable.
- Homogeneity of variance/covariance (homoscedasticity): Variances among group variables are the same across levels of predictors. Can be tested with Box's M statistic.
- It has been suggested, however, that linear discriminant analysis be used when covariances are equal, and that quadratic discriminant analysis may be used when covariances are not equal.
- Multicollinearity: Predictive power can decrease with an increased correlation between predictor variables.
- Independence: Participants are assumed to be randomly sampled, and a participant’s score on one variable is assumed to be independent of scores on that variable for all other participants.
- It has been suggested that discriminant analysis is relatively robust to slight violations of these assumptions, and it has also been shown that discriminant analysis may still be reliable when using dichotomous variables (where multivariate normality is often violated).

Discriminant analysis works by creating one or more linear combinations of predictors, creating a new variable for each function. These functions are called discriminant functions. The number of functions possible is either Ng-1 where Ng = number of groups, or p (the number of predictors), whichever is smaller. The first function created maximizes the differences between groups on that function. The second function maximizes differences on that function, but also must not be correlated with the previous function. This continues with subsequent functions with the requirement that the new function not be correlated with any of the previous functions.

Given group j, with $\mathbb{R}_j$ sets of sample space, there is a discriminant rule such that if $x \in \mathbb{R}_j$, then $x \in j$. Discriminant analysis then, finds "good" regions of $\mathbb{R}_j$ to minimize classification error, therefore leading to a high percent correct classified in the classification table. Each function is given a discriminant score to determine how well it predicts group placement.

Structure Correlation Coefficients: The correlation between each predictor and the discriminant score of each function. This is a whole correlation.

- Standardized Coefficients: Each predictor's unique contribution to each function, therefore this is a partial correlation. Indicates the relative importance of each predictor in predicting group assignment from each function.
- Functions at Group Centroids: Mean discriminant scores for each grouping variable are given for each function. The farther apart the means are, the less error there will be in classification.

Discrimination rules

- Maximum likelihood: Assigns x to the group that maximizes population (group) density.
- Bayes Discriminant Rule: Assigns x to the group that maximizes $\pi_i f_i(x)$, where $f_i(x)$ represents the prior probability of that classification, and π_i represents the population density.
- Fisher's linear discriminant rule: Maximizes the ratio between SS_{between} and SS_{within} , and finds a linear combination of the predictors to predict group.

Eigen values

An eigen value in discriminant analysis is the characteristic root of each function. It is an indication of how well that function differentiates the groups, where the larger the eigenvalue, the better the function differentiates. This however, should be interpreted with caution, as eigenvalues have no upper limit. The eigenvalue can be viewed as a ratio of $SS_{\text{between}}$and SS_{within} as in ANOVA when the dependent variable is the discriminant function, and the groups are the levels of the IV. This means that the largest eigenvalue is associated with the first function, the second largest with the second, etc.

Effect size

Some suggest the use of eigenvalues as effect size measures, however, this is generally not supported. Instead, the canonical correlation is the preferred measure of effect size. It is similar to the eigenvalue, but is the square root of the ratio of SS_{between} and SS_{total}. It is the correlation between groups and the function. Another popular measure of effect size is the percent of variance for each function. This is calculated by: ($\lambda_x/\Sigma\lambda_i$) X 100 where λ_x is the eigenvalue for the function and $\Sigma\lambda_i$ is the sum of all eigenvalues. This tells us how strong the prediction is for that particular function compared to the others. Percent correctly classified can also be analyzed as an effect size. The kappa value can describe this while correcting for chance agreement.

Variations

- Multiple discriminant analysis (MDA): related to MANOVA. Has more than two groups, and uses multiple dummy variables.
- Sequential discriminant analysis: assesses the importance of a set of IVs over and above a set of controls. In this case, the controls are entered first, and then the IVs.
- Stepwise discriminant analysis: Selects the most correlated predictor first, removes that variance in the grouping variable then adds the next most correlated and continues

until the change in canonical correlation is not significant. Of course, both forward and backward stepwise procedures may be performed.

In DFA one wishes to predict group membership from a set of (usually continuous) predictor variables. In the most simple case one has two groups and *p* predictor variables. A linear discriminant equation, $D_i = a + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$, is constructed such that the two groups differ as much as possible on *D*. That is, the weights are chosen so that were you to compute a discriminant score (D_i) for each subject and then do an ANOVA on *D*, the ratio of the between groups sum of squares to the within groups sum of squares is as large as possible. The value of this ratio is the eigenvalue. “Eigen” can be translated from the German as “own,” “peculiar,” “original,” “singular,” etc.

The eigenvalue $= \frac{SS_{between\_groups}}{SS_{within\_groups}}$ on *D* (the quantity maximized by the discriminant function coefficients obtained). The canonical correlation $= \sqrt{\frac{SS_{between\_groups}}{SS_{total}}}$ on *D* (equivalent to *eta* in an ANOVA and equal to the point biserial *r* between Group and *D*),

Wilks lambda is used to test the null hypothesis that the populations have identical means on *D*. Wilks lambda is $\Lambda = \frac{SS_{within\_groups}}{SS_{total}}$, so the smaller the Λ the more doubt cast upon that null hypothesis. SPSS uses a χ^2 approximation to obtain a significance level. We can determine how much of the variance in the grouping variable is explained by our predictor variables by subtracting the Λ from one.

DFA is mathematically equivalent to a MANOVA. Looking at Λ from the perspective of a MANOVA, when we combine the rating scales with weights that maximize group differences on the resulting linear combination, the groups do differ significantly from one another. Such a MANOVA is sometimes done prior to doing univariate analyses to provide a bit of protection against inflation of alpha. Recall that the grouping variable is predictor variable in MANOVA (is it what is being predicted in DFA) and the rating scales are the MANOVA outcome variables (and our DFA predictor variables). If the MANOVA is not significant, we stop. If it is significant, we may go on to do an ANOVA on each dependent variable. SPSS gave us those ANOVAs.

We have created (or discovered) a dimension (like a component in PCA) on which the two groups differ. The univariate ANOVAs may help us explain the nature of the relationship between this discriminant dimension and the grouping variable. For example, some of the variates may have a significant relationship with the grouping variable and others might not, but the univariate ANOVAs totally ignore the correlations among the variates. It is possible for the groups to differ significantly on *D* but not on any one predictor by itself.

The standardized discriminant function coefficients may help. These may be treated as *Beta* weights in a multiple regression predicting *D* from *z*-scores on the X’s, $D_i = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p$. Of course, one must realize that these coefficients reflect the contribution of one variate in the context of the other variates in the model. A low standardized coefficient might mean that the groups do not differ much on that variate or it might just mean that that variate’s correlation with the grouping variable is redundant with that of another variate in the model. Suppressor effects can also occur.

Correlations between variates and *D* may also be helpful. These are available in the loading or structure matrix. Generally, any variate with a loading of .30 or more is considered to be important in defining the discriminant dimension. These correlations may help us understand the discriminant function we have created.

If your primary purpose is to predict group membership from the variates (rather than to examine group differences on the variates), you need to do classification. SPSS classifies subjects into predicted groups using Bayes' rule: $p(G_i \mid D) = \dfrac{p(G_i) \times p(D \mid G_i)}{\sum_{i=1}^{g} p(G_i) \times p(D \mid G_i)}$.

Each subject's discriminant score is used to determine the posterior probabilities of being in each of the two groups. The subject is then classified (predicted) to be in the group with the higher posterior probability.

By default, SPSS assumes that all groups have equal prior probabilities. For two groups, each prior = ½, for three, 1/3, etc. I asked SPSS to use the group relative frequencies as priors, which should result in better classification.

Another way to classify subjects is to use Fisher's classification function coefficients. For each subject a *D* is computed for each group and the subject classified into the group for which e's *D* is highest. To compute a subjects D_1 you would multiply e's scores on the 22 rating scales by the indicated coefficients and sum them and the constant. For e's D_2 you would do the same with the coefficients for Group 2. If $D_1 > D_2$ then you classify the subject into Group 1, if $D_2 > D_1$, the you classify em into Group 2.

For validity of significance tests, one generally does not worry about this if sample sizes are equal, and with unequal sample sizes one need not worry unless the $p < .001$. The DFA is thought to be very robust and Box's *M* is very sensitive. Non-normality also tends to lower the *p* for Box's M. The classification procedures are not, however, so robust as the significance tests are. One may need to transform variables or do a quadratic DFA (SPSS won't do this) or ask that separate rather than pooled variance-covariance matrices be used. Pillai's criterion (rather than Wilk's Λ) may provide additional robustness for significance testing -- although not available with SPSS discriminant, this criterion is available with SPSS MANOVA.

ANOVA on *D*. Conduct an ANOVA comparing the verdict groups on the discriminant function. Then you can demonstrate that the DFA eigenvalue is equal to the ratio of the SS_{between} to SS_{within} from that ANOVA and that the ratio of SS_{between} to SS_{total} is the squared canonical correlation coefficient from the DFA.

Comparison and validation of models

R^2 (Coefficient of Determination)

It is in general used for checking the adequacy of the model. R^2 is given by the following formula

$$R^2 = 1 - \frac{ss_{res}}{ss_t}$$

where ss_{res} and ss_t are the residual sum of square and the total sum of square respectively.

R^2 never decreases when a regressor is added to the model, regardless of the value of the contribution of the variable to the model. Therefore, it is difficult to judge whether an increase in R^2 is really telling anything important. So it is preferable to use Adjusted R^2 when models to be compared are based on different number of regressors. Adjusted R^2 is given by the following formula

$$R_{adj}{}^2 = 1 - \frac{ss_{res}/(n-p)}{ss_t/(n-1)}$$

where $ss_{res}/(n\text{-}p)$ is the residual mean square and $ss_t/(n\text{-}1)$ is the total mean square. The total mean square is constant regardless of how many variables are in the model. On adding a regressor in the model Adjusted R^2 increases only if the addition of the regressor reduces the residual mean square. It also penalizes for adding terms that are not helpful, so it is very important in evaluating and comparing the candidate regression models.

Percent Deviation

This measures the deviation (in percentage) of forecast from the actual yield data. The formula for calculating the percent deviation of forecast is given below

$$percentage\, deviation = \frac{(actual\ yield - forecasted\ yield)}{actual\ yield} \times 100$$

Root Mean Square Error (RMSE)

It is also a measure of comparing two models. The formula of RMSE is given bellow

$$RMSE = [\{\frac{1}{n}\sum_{i=1}^{n}(O_i - E_i)^2\}]^{\frac{1}{2}}$$

O_i and the E_i are the observed and forecasted value of the crop yield respectively and n is the number of years for which forecasting has been done.

PRINCIPAL COMPONENTS ANALYSIS

Multivariate data consist of observations on several different variables for a number of individuals or subjects. Data of this type arise in all the branches of science, ranging from psychology to biology, and methods of analyzing multivariate data constitute an increasingly important area of statistics. Indeed, the vast majority of data in forestry is multivariate and proper handling of such data is highly essential. Principal components analysis (PCA) and Factor analysis (FA) are multivariate techniques applied to a single set of variables to discover which sets of variables in the set form coherent subsets that are relatively independent of one another. The details of PCA and FA are discussed as below.

Principal Components Analysis

Most of the times the variables under study are highly correlated and as such they are effectively "saying the same thing". To examine the relationships among a set of *p* correlated variables, it may be useful to transform the original set of variables to a new set of uncorrelated variables called *principal components*. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data.

Let $x_1, x_2, x_3, \dots, x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11}\,x_1 + a_{12}\,x_2 + \dots\dots + a_{1p}\,x_p$$

such that variance of z_1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + \dots\dots + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then $\mathrm{Var}(z_1)$ can be increased simply by multiplying any a_{1j}s by a constant factor

The second principal component is defined as

$$z_2 = a_{21}\,x_1 + a_{22}\,x_2 + \dots\dots + a_{2p}\,x_p$$

such that $\mathrm{Var}(z_2)$ is as large as possible next to $\mathrm{Var}(\ z_1\ )$subject to the constraint that

$$a_{21}^2 + a_{22}^2 + \dots\dots + a_{2p}^2 = 1 \quad \text{and} \quad \mathrm{cov}(z_1, z_2) = 0 \text{ and so on.}$$

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus, reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, Principal Component Analysis is more of a means to an end rather than an end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique which does not require user to specify the statistical model or assumption about distribution of original variates. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Component Analysis transforms original set of variables to new set of uncorrelated variables, it is worth stressing that if original variables are uncorrelated, then there is no point in carrying out principal component analysis.

Computation of principal components :

Let us consider the following data on average minimum temperature (x_1), average relative humidity at 8 hrs. (x_2), average relative humidity at 14 hrs. (x_3) and total rainfall in cm. (x_4) pertaining to Raipur district from 1970 to 1986 for kharif season from 21st May to 7th Oct.

	X_1	x_2	x_3	x_4
	25.0	86	66	186.49
	24.9	84	66	124.34
	25.4	77	55	98.79
	24.4	82	62	118.88
	22.9	79	53	71.88
	7.7	86	60	111.96
	25.1	82	58	99.74
	24.9	83	63	115.20
	24.9	82	63	100.16
	24.9	78	56	62.38
	24.3	85	67	154.40
	24.6	79	61	112.71
	24.3	81	58	79.63
	24.6	81	61	125.59
	24.1	85	64	99.87
	24.5	84	63	143.56
	24.0	81	61	114.97
Mean	23.56	82.06	61.00	112.97
S.D.	4.13	2.75	3.97	30.06

with the variance co-variance matrix.

$$\Sigma = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{bmatrix}$$

Find the eigen values and eigen vectors of the above matrix. Arrange the eigen values in decreasing order. Let the eigen values in decreasing order and corresponding eigen vectors are

$\lambda_1 = 916.902 \quad a_1 = (0.006, \; 0.061, \; 0.103, \; 0.993)$

$\lambda_2 = 18.375 \quad a_2 = (0.955, \; -0.296, \; 0.011, \; 0.012)$

$\lambda_3 = 7.87 \quad a_3 = (0.141, \; 0.485, \; 0.855, \; -0.119)$

$\lambda_4 = 1.056 \quad a_4 = (0.260, \; 0.820, \; -0.509, \; 0.001)$

The principal components for this data will be

$$\begin{aligned} z_1 &= 0.006\ x_1 + 0.061\ x_2 + 0.103\ x_3 + 0.993\ x_4 \\ z_2 &= 0.955\ x_1 - 0.296\ x_2 + 0.011\ x_3 + 0.012\ x_4 \\ z_3 &= 0.141\ x_1 + 0.485\ x_2 + 0.855\ x_3 - 0.119\ x_4 \\ z_4 &= 0.26\ x_1 + 0.82\ x_2 - 0.509\ x_3 + 0.001\ x_4 \end{aligned}$$

The variance of principal components will be eigen values i.e.

Var(z_1) = 916.902, Var(z_2) = 18.375, Var (z_3) = 7.87, Var(z_4) = 1.056

The total variation explained by original variables is

$$= \text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3) + \text{Var}(x_4)$$

$$= 17.02 + 7.56 + 15.75 + 903.87 = 944.20$$

The total variation explained by principal components is

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = .97$$

Continuing, the first two components account for a proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = .99$$

of the total variance.

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of x_i s in equations of z_i s. For above data, the first two principal components for first observation i.e. for year 1970 can be worked out as

$$z_1 = 0.006 \text{ x } 25.0 + 0.061 \text{ x } 86 + 0.103 \text{ x } 66 + 0.993 \text{ x } 186.49 = 197.380$$
$$z_2 = 0.955 \text{ x } 25.0 - 0.296 \text{ x } 86 + 0.011 \text{ x } 66 + 0.012 \text{ x } 186.49 = 1.383$$

Similarly for the year 1971

$$z_1 = 0.006 \text{ x } 24.9 + 0.061 \text{ x } 84 + 0.103 \text{ x } 66 + 0.993 \text{ x } 124.34 = 135.54$$
$$z_2 = 0.955 \text{ x } 24.9 - 0.296 \text{ x } 84 + 0.011 \text{ x } 66 + 0.012 \text{ x } 124.34 = 1.134$$

Thus the whole data with four variables can be converted to a new data set with two principal components.

Note: The principal components depend on the scale of measurement, for example, if in the above example X_1 is measured in ^{0}F instead of ^{0}C and X_4 in mm in place of cm, the data gives different principal components when transformed to original x's. In very specific situations results are same. The conventional way of getting around this problem is to use standardized variables with unit variances, i.e., correlation matrix in place of dispersion matrix. But the principal components obtained from original variables as such and from correlation matrix will not be same and they may not explain the same proportion of variance in the system. Further more, one set of principal components is not simple function of the other. When the variables are standardized, the resulting variables contribute almost equally to the principal components determined from correlation matrix. Variables should probably be standardized if they are measured on scales with widely differing ranges or if measured units are not commensurate. Often population dispersion matrix or correlation matrix are not available. In such situations sample dispersion matrix or correlation matrix can be used.

Applications of principal components:

- The most important use of principal component analysis is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then first few components' scores can be utilized in subsequent analysis in place of original variables.
- Plotting of data becomes difficult with more than three variables. Through principal component analysis, it is often possible to account for most of the variability in the data by first two components, and it is possible to plot the values of first two components scores for each individual. Thus, principal component analysis enables us to plot the data in two dimensions. Particularly detection of outliers or clustering of individuals will be easier through this technique. Often, use of principal component analysis reveals grouping of variables which would not be found by other means.
- Reduction in dimensionality can also help in analysis where no. of variables is more than the number of observations, for example, in discriminant analysis and regression analysis. In such cases, principal component analysis is helpful by reducing the dimensionality of data.
- Multiple regression can be dangerous if independent variables are highly correlated. Principal component analysis is the most practical technique to solve the problem. Regression analysis can be carried out using principal components as regressors in place of original variables. This is known as principal component regression.

FACTOR ANALYSIS

Factor analysis has originated in the field of psychology to define the concepts like intelligence, attitude, etc. The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called *factors*. Under the factor model assuming linearity, each response variate is represented as a linear function of a small number of unobservable common factors and a single latent *specific factor*. The common factors generate the covariances among the observable responses while the specific terms contribute only to the variances of their particular response. Basically the factor model is motivated by the following argument - Suppose variables can be grouped by their correlations, i.e., all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. For example, for an individual, marks in different subjects may be governed by aptitudes (common factors) and individual variations (specific factors) and interest may lie in obtaining scores on unobservable aptitudes (common factors) from observable data on marks in different subjects.

The Factor Model

Suppose observations are made on p variables for n individuals (x_{ij}, i=1,2,…p; j=1,2,…n). The factor analysis model assumes that there are m underlying factors (m<p) and each observed variable is a linear function of these factors and specific factor, so that

$$x_{ij} = a_{i1} f_{1j} + a_{i2} f_{2j} + \ldots\ldots\ldots\ldots + a_{im} f_{mj} + a_{i0} y_{ij} \qquad j = 1,2,\ldots,p$$

where a_{i1}, a_{i2}, ……., a_{im} are *factor loadings* given to i-th variable corresponding to m common hypothetical factors of j-th respondent (f_{1j}, f_{2j}, ….., f_{mj}) and a_{i0} is the loading given to factor specific to i-th variable pertaining to j-th respondent (y_{ij}).

The proportion of the variance of the j-th variable contributed by the *m* common factors is called the *j*-th *communality* and the proportion due to the specific factors is called the *uniqueness*, or *specific variance.*

Factor analysis involves :

- Deciding number of common factors (m)
- Estimating factor loadings (a_{ik})
- Calculating factor scores (f_{kj})

Methods of Estimation

Factor analysis is done in two parts, first solution is obtained by placing some restrictions and then final solution is obtained by rotating this solution. There are two most popular methods available in literature for parameter estimation, the *principal component* (and the related *principal factor*) method and the *maximum likelihood method*. The solution from either method can be rotated in order to simplify the interpretation of factors i.e. either factor loadings are close to unity or close to zero. The most popular method for orthogonal rotation is *Varimax Rotation method*. In some specific situations, oblique rotations are also used. It is always prudent to try more than one method of solution. If the factor model is appropriate for the problem at hand, the solutions should be consistent with one another. The estimation and rotation methods require iterative calculations that must be done on a computer. If variables are uncorrelated factor analysis will not be useful. In these circumstances, the specific factors play the dominant role, whereas the major aim of the factor analysis is to determine a few important common factors.

Number of factors is theoretically given by rank of population variance covariance matrix. However, in practice, number of common factors retained in the model is increased until a suitable proportion of total variance is explained. Another convention, frequently encountered in packaged computer programs is to set m equal to the number of eigenvalues greater than one (for example, in SAS and SPSS).

As in principal component analysis, principal factor method for factor analysis depends upon unit of measurements. If units are changed, the solution will change. However, in this approach estimated factor loadings for a given factor do not change as the number of factors is increased. In contrast to this, in maximum likelihood method, the solution does not change if units of measurements are changed. However, in this method the solution changes if number of common factors is changed.

Example: In a consumer - preference study, a random sample of customers were asked to rate several attributes of a new product. The response on a 5-point semantic differential scale were tabulated and the attribute correlation matrix constructed which is given below:

Attribute		**Correlation matrix**				
		1	**2**	**3**	**4**	**5**
Taste	**1**	1	.02	.96	.42	.01
Good buy for money	**2**	.02	1	.13	.71	.85
Flavor	**3**	.96	.13	1	.5	.11
Suitable for snack	**4**	.42	.71	.50	1	.79
Provides energy	**5**	.01	.85	.11	.79	1

It is clear from the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is “closer” to the (2,5) group than (1,3) group. Observing the results, one can expect that the apparent linear relationships between the variables can be explained in terms of, at most, two or three common factors.

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE
Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	1	2	3	4	5
Eigenvalue	2.8531	1.8063	0.2045	0.1024	0.0337
Difference	1.0468	1.6018	0.1021	0.0687	
Proportion	0.5706	0.3613	0.0409	0.0205	0.0067
Cumulative	0.5706	0.9319	0.9728	0.9933	1.0000

```
                  FACTOR1          FACTOR2
        TASTE     0.55986          0.81610
        MONEY     0.77726          -0.52420
        FLAVOR    0.64534          0.74795
        SNACK     0.93911          -0.10492
        ENERGY    0.79821          -0.54323

      Variance explained by each factor
          FACTOR1        FACTOR2
         2.853090        1.806332

    Final Communality Estimates: Total = 4.659423

      TASTE     MONEY    FLAVOR     SNACK    ENERGY
    0.979461  0.878920  0.975883  0.892928  0.932231

   Residual Correlations with Uniqueness on the Diagonal

             TASTE       MONEY       FLAVOR      SNACK      ENERGY
   TASTE     0.02054      0.01264    -0.01170    -0.02015     0.00644
   MONEY      0.01264     0.12108     0.02048    -0.07493    -0.05518
   FLAVOR   -0.01170      0.02048     0.02412    -0.02757     0.00119
   SNACK    -0.02015     -0.07493    -0.02757     0.10707     -.01660
   ENERGY    0.00644     -0.05518     0.00119    -0.01660     0.06777

Rotation Method: Varimax

 Rotated Factor Pattern
                  FACTOR1          FACTOR2
        TASTE     0.01970          0.98948
        MONEY     0.93744          -0.01123
        FLAVOR    0.12856          0.97947
        SNACK     0.84244          0.42805
        ENERGY    0.96539          -0.01563

      Variance explained by each factor

          FACTOR1        FACTOR2
         2.537396        2.122027

    Final Communality Estimates: Total = 4.659423

      TASTE     MONEY    FLAVOR     SNACK    ENERGY

     0.979461  0.878920  0.975883  0.892928  0.932231

Initial Factor Method: Maximum Likelihood
Eigenvalues of the Weighted Reduced Correlation Matrix:
Total = 84.5563187
```

```
Average = 16.9112637
                         1          2          3              4          5
          Eigenvalue     59.7487   24.8076    0.1532     -0.0025    -0.1507
          Difference     34.9411   24.6544    0.1557      0.1482
          Proportion     0.7066    0.2934     0.0018     -0.0000    -0.0018
          Cumulative      0.7066    1.0000     1.0018     1.0018    1.0000

                                   Factor Pattern
                              FACTOR1              FACTOR2
                     TASTE    0.97601              -0.13867
                     MONEY    0.14984               0.86043
                     FLAVOR   0.97908              -0.03180
                     SNACK    0.53501               0.73855
                     ENERGY   0.14567               0.96257

 Variance explained by each factor
                              FACTOR1     FACTOR2
                    Weighted  59.748704   24.807616
                  Unweighted  2.241100    2.232582

                          TASTE     MONEY     FLAVOR    SNACK     ENERGY
           Communality  0.971832  0.762795  0.959603  0.831686  0.947767

Rotation Method: Varimax
Rotated Factor Pattern

                           FACTOR1          FACTOR2
                     TASTE   0.02698         0.98545
                     MONEY   0.87337         0.00342
                     FLAVOR  0.13285          0.97054
                     SNACK   0.81781         0.40357
                     ENERGY  0.97337         -0.01782

                    Variance explained by each factor
                              FACTOR1     FACTOR2
                    Weighted  25.790361   58.765959
                  Unweighted  2.397426    2.076256

                    TASTE     MONEY     FLAVOR    SNACK     ENERGY
       Communality  0.971832  0.762795  0.959603  0.831686  0.947767
```

It can be seen that two factor model with factor loadings shown above is providing a good fit to the data as the first two factors explains 93.2% of the total standardized sample variance, i.e., $\left(\frac{\lambda_1+\lambda_2}{p}\right)x100$, where p is the number of variables. It can also be seen from the results that there is no clear-cut distinction between factor loadings for the two factors before rotation but after rotation the same is clear.

DATA DIAGNOSTICS AND TRANSFORMATION

The raw data consist of measurements of some attribute on a collection of individuals. The measurement would have been made in one of the following scales *viz*., nominal, ordinal, interval or ratio scale.

Levels of Measurement

- **Nominal scale** refers to measurement at its weakest level when number or other symbols are used simply to classify an object, person or characteristic, *e.g.*, state of health (healthy, diseased).
- **Ordinal scale** is one wherein given a group of equivalence classes, the relation greater than holds for all pairs of classes so that a complete rank ordering of classes is possible, *e.g*., socio-economic status.
- When a scale has all the characteristics of an ordinal scale, and when in addition, the distances between any two numbers on the scale are of known size, **interval scale** is achieved, e.*g*., temperature scales like centigrade or Fahrenheit.
- An interval scale with a true zero point as its origin forms a ratio scale. In a **ratio scale**, the ratio of any two scale points is independent of the unit of measurement, e.g., height of trees.

The data can be classified as qualitative/quantitative depending on the levels based on which the observations are collected. There are several statistical procedures available in literature for the analysis of data which are broadly classified in to two categories viz., parametric tests and non-parametric tests. A parametric test specifies certain conditions about the distribution of responses in the population from which the research sample was drawn. The meaningfulness of the results of a parametric test depends on the validity of these assumptions. A nonparametric test is based on a model that specifies very general conditions and none regarding the specific form of the distribution from which the sample was drawn. Hence nonparametric tests are also known as distribution free tests. Certain assumptions are associated with most nonparametric statistical tests, but these are fewer and weaker than those of parametric tests.

Nonparametric test statistics utilize some simple aspects of sample data such as the signs of measurements, order relationships or category frequencies. Therefore, stretching or compressing the scale does not alter them. As a consequence, the null distribution of the nonparametric test statistic can be determined without regard to the shape of the parent population distribution. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and, as a result, if the data have one or two outliers, their influence is negated.

Besides, the interpretation of data based on analysis of variance (ANOVA) is valid only when the following assumptions are satisfied:

1. **Additive Effects:** Treatment effects and block (environmental) effects are additive.
2. **Independence of errors:** Experimental errors are independent.
3. **Homogeneity of Variances:** Errors have common variance.
4. **Normal Distribution:** Errors follow a normal distribution.

Also the statistical tests t, F, z, etc. are valid under the assumption of independence of errors and normality of errors. The departures from these assumptions make the interpretation based on these statistical techniques invalid. Therefore, it is necessary to detect the deviations and apply the appropriate remedial measures.

- The assumption of independence of errors, *i.e.*, error of an observation is not related to or depends upon that of another. This assumption is usually assured with the use of proper randomization procedure. However, if there is any systematic pattern in the arrangement of treatments from one replication to another, errors may be non-independent. This may be handled by using nearest neighbour methods in the analysis of experimental data.
- The assumption of additive effects can be defined and detected in the following manner:

Additive Effects

The effects of two factors, say, treatment and replication, are said to be additive if the effect of one-factor remains constant over all the levels of other factors. A hypothetical set of data from a randomized complete block (RCB) design, with 2 treatments and 2 replications, with additive effects is given in Table 1.

Table 1

Treatment	**Replication**		**Replication Effect**
	I	**II**	**I - II**
A	190	125	*65*
B	170	105	*65*
Treatment Effect (A-B)	*20*	*20*	

Here, the treatment effect is equal to 20 for both replications and replication effect is 65 for both treatments.

When the effect of one factor is not constant at all the levels of other factor, the effects are said to be non-additive. A common departure from the assumption of additivity in biological experiments is one where the effects are multiplicative. Two factors are said to have multiplicative effects if their effects are additive only when expressed in terms of percentages. Table 2 illustrates a hypothetical set of data with multiplicative effects.

Table 2

Treatment	**Replication**		**Replication Effect**	
	I	**II**	**I - II**	**100(I - II)/II**
A	200 (2.30103)	125 (2.09691)	*75* (0.20412)	*60*
B	160 (2.20412)	100 (2.0000)	*60* (0.20412)	*60*
Treatment Effect (A-B)	*40* (0.09691)	*25* (0.09691)		
100 (A - B)/B	*25*	*25*		

In this case, the treatment effect is not constant over replications and the replication effect is not constant over treatments. However, when both treatment effect and replication effect are expressed in terms of percentages, an entirely different pattern emerges. For such violations of assumptions, Logarithmic transformation is quite suitable. For illustration, the Logarithmic transformation of data in Table 2 is given in brackets.

This is, however a crude method for testing the additivity. Tukey (1949) gave a statistical test for testing the additivity in a RCB design. This test is known as one degree of freedom test for non-additivity. In this test, one degree of freedom is isolated from error and this degree of freedom is called as the degree of freedom for non-additivity. In the sequel, we describe the procedure in brief.

Suppose that an experiment has been conducted to compare *v* treatments using RCB design with *r* replications. Let y_{ij} denote the observed value of the response variable for i^{th} treatment in j^{th} replication; *i* = 1,2, …, *v*; *j* = 1,2, …, *r*. Arrange the data in a *v* × *r* table as given below:

Treatment	1	2	$\cdots$	*j*	$\cdots$	*r*	Treatment Total	Treatment Mean	Deviations from Grand Mean	Sum of Cross Product
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1r}	$T_{1.}$	$\bar{y}_{1.}$	$d_{1.}$	C_1
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2r}	$T_{2.}$	$\bar{y}_{2.}$	$d_{2.}$	C_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ir}	$T_{i.}$	$\bar{y}_{i.}$	$d_{i.}$	C_i
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
v	y_{v1}	y_{v2}	$\cdots$	y_{vj}	$\cdots$	y_{vr}	$T_{v.}$	$\bar{y}_{v.}$	$d_{v.}$	C_v
Replication Total	$R_{.1}$	$R_{.2}$	$\cdots$	$R_{.j}$	$\cdots$	$R_{.r}$	*G* (Grand total)			
Replication Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\cdots$	$\bar{y}_{.j}$	$\cdots$	$\bar{y}_{.r}$		$GM = \dfrac{G}{vr}$		
Deviation from Grand Mean	$d_{.1}$	$d_{.2}$	$\cdots$	$d_{.j}$	$\cdots$	$d_{.r}$				

where $T_{i.} = \sum_{j=1}^{r} y_{ij}$; $\bar{y}_{i.} = T_{i.} / r$; $R_{.j} = \sum_{i=1}^{v} y_{ij}$; $\bar{y}_{.j} = R_{.j} / v$; $d_{i.} = \bar{y}_{i.} - GM$

$d_{.j} = \bar{y}_{.j} - GM$; $C_i = \sum_{j=1}^{r} y_{ij} \times d_{.j}$

Obtain $L = \sum_{i=1}^{v} C_i d_{i.}$; $D_1 = \sum_{i=1}^{v} d_{i.}^2$; $D_2 = \sum_{j=1}^{r} d_{.j}^2$

Sum of squares due to non-additivity (SSNA)= $\dfrac{L^2}{D_1 \times D_2}$

The sum of squares due to treatments, replications and total sum of squares are given by

Sum of squares due to treatments (SST) = $\sum_{i=1}^{v} \dfrac{T_{i.}^2}{r} - \dfrac{G^2}{vr}$

Sum of squares due to replications (SSR) = $\sum_{j=1}^{r} \dfrac{R_{.j}^2}{v} - \dfrac{G^2}{vr}$

Total sum of squares (TSS) = $\sum_{i=1}^{v}\sum_{j}^{r} y_{ij}^2 - \frac{G^2}{vr}$

Sum of squares due to Error (SSE) = TSS – SST-SSR-SSNA

Then the outline of ANOVA table is

Source	**df**	**SS**	**MS**
Treatments	v-1	SST	MST
Replications	r-1	SSR	MSR
Non-additivity	1	SSNA	MSNA
Error	(v-1)(r-1)-1	SSE	MSE
Total	***vr*-1**	**TSS**	

The mean squares (MS) are obtained by dividing sum of squares (SS) by corresponding degrees of freedom (df). The non-additivity is tested by F-statistic with 1 and (v-1)(r-1)-1 degree of freedom calculated value of F = $\frac{\text{MSNA}}{\text{MSE}}$.

Normality of Errors

The assumptions of homogeneity of variances and normality are generally violated together. To test the validity of normality of errors for the character under study, one can take help of Normal Probability Plot, Anderson-Darling Test, D'Augstino's Test, Shapiro - Wilk's Test, Ryan-Joiner test, Kolmogrov-Smirnov test, etc. In general moderate departures from normality are of little concern in the fixed effects ANOVA as F - test is slightly affected but in case of random effects, it is more severely impacted by non-normality. The significant deviations of errors from normality, makes the inferences invalid. So before analyzing the data, it is necessary to convert the data to a scale that it follows a normal distribution. In the data from designed field experiments, we do not directly use the original data for testing of normality or homogeneity of observations because this is embedded with the treatment effects and some of other effects like block, row, column, etc. So there is need to eliminate these effects from the data before testing the assumptions of normality and homogeneity of variances. For eliminating the treatment effects and other effects we fit the model corresponding to the design adopted and estimate the residuals. These residuals are then used for testing the normality of the observations. In other words, we want to test the null hypothesis H_0: errors are normally distributed against alternative hypothesis H_1: errors are not normally distributed. For details on these tests one may refer to D'Agostino and Stephens (1986). Most of the standard statistical packages available in the market are capable of testing the normality of the data. In SAS and SPSS commonly used tests are Shapiro-Wilk test and Kolmogrov-Smirnov test. MINITAB uses three tests viz. Anderson-Darling, Ryan-Joiner, Kolmogrov-Smirnov for testing the normality of data.

Homogeneity of Error Variances

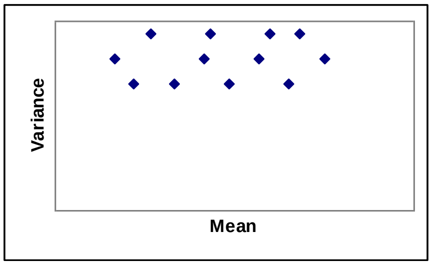
A crude method for detecting the heterogeneity of variances is based on scatter plots of means and variance or range of observations or errors, residual vs fitted values, etc. To be clearer, let Y_{ij} be the observation pertaining to i^{th} treatment $\left(i = 1(1)v\right)$ in the j^{th} replication $\left(j = 1(1)r_i\right)$. Compute the mean and variance for each treatment across the replications (the range can be used in place of variance) as

$$\text{Mean} = \bar{Y}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij} \; ; \qquad \text{Variance} = S_i^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

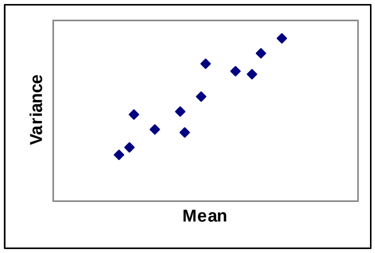
Draw the scatter plot of mean vs variance (or range). If $S_{i.}^2$'s $\left(i = 1(1)v\right)$ are equal (constant) or nearly equal, then the variances are homogeneous. Based on these scatter plots, the heterogeneity of variances can be classified into two types:

1. Where the variance is functionally related to mean.
2. Where there is no functional relationship between the variance and the mean.

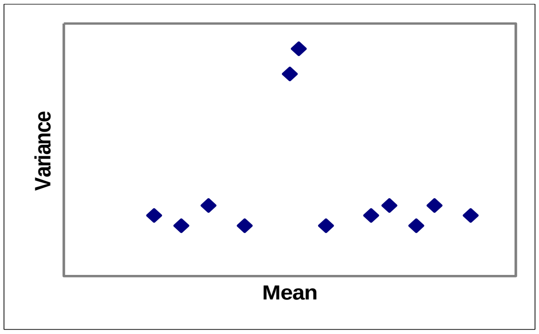
For illustration some scatter - diagrams of mean and variances (or range) are given as:



(a) Homogeneous variance



(b) Heterogeneous variance where variance is proportional to mean



(c) Heterogeneous variance without any functional relationship between variance and mean

The first kind of variance heterogeneity (figure b) is usually associated with the data whose distribution is non-normal *viz.*, negative binomial, Poisson, binomial, etc. The second kind of variance heterogeneity usually occurs in experiments, where, due to the nature of treatments tested, some treatments have errors that are substantially higher (lower) than others. For example, in varietal trials, where various types of breeding material are being compared, the size of variance between plots of a particular variety will depend on the degree of genetic homogeneity of material being tested. The variance of F_2 generation, for example, can be expected to be higher than that of F_1 generation because genetic variability in F_2 is much higher than that in F_1. The variances of varieties that are highly tolerant of or highly susceptible to, the stress being tested are expected to be smaller than those of having moderate degree of tolerance. Also in testing yield response to a chemical treatment, such as, fertilizer, insecticide or herbicide, the non-uniform application of chemical treatments may result in a higher variability in the treated plots than that in the untreated plots.

The scatter-diagram of means and variances of observations for each treatment across the replications gives only a preliminary idea about homogeneity of error variances. Statistically the homogeneity of error variances is tested using Bartlett's test for normally distributed errors and Levene test for non-normal errors. These tests are described in the sequel.

Bartlett's Test for Homogeneity of Variances

Let there are *v*- independent samples drawn from same population and *i*th sample is of size r_i and $(r_1 + r_2 + ... + r_v) = N$. In the present case, the independent samples are the residuals of the observations pertaining to *v* treatments and *i*th sample size is the number of replications of the treatment *i*. One wants to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_v^2$ against the alternative hypothesis H_1 : at least two of the σ_i^2's are not equal, where σ_i^2 is the error variance for treatment *i*.

Let e_{ij} denotes the residual pertaining to the observation of treatment *i* from replication *j*, then it can easily be shown that the sum of residuals pertaining to a given treatment is zero. In this test $S_i^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}(e_{ij} - \bar{e}_{i.})^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i} e_{ij}^2$ is taken as unbiased estimate of σ_i^2. The procedure involves computing a statistic whose sampling distribution is closely approximated by the χ^2 distribution with *v - 1* degrees of freedom. The test statistic is

$$\chi_0^2 = 2.3026\frac{q}{c}$$

and null hypothesis is rejected when $\chi_0^2 > \chi_{\alpha,v-1}^2$, where $\chi_{\alpha,v-1}^2$ is the upper α percentage point of χ^2 distribution with *v - 1* degrees of freedom.

To compute χ_0^2, follow the steps:

Step 1: Compute mean and variance of all *v*-samples.

Step 2: Obtain pooled variance $S_p^2 = \frac{\sum_{i=1}^{v}(r_i - 1)S_i^2}{N - v}$

Step 3: Compute $q = (N - v)\log_{10} S_p^2 - \sum_{i=1}^{v}(r_i - 1)\log_{10} S_i^2$

Step 4: Compute $c = 1 + \frac{1}{3(v-1)}\left(\sum_{i=1}^{v}(r_i - 1)^{-1} - (N - v)^{-1}\right)$

Step 5: Compute χ_0^2.

Bartlett's χ^2 test for homogeneity of variances is a modification of the normal-theory likelihood ratio test. While Bartlett's test has accurate Type I error rates and optimal power when the underlying distribution of the data is normal, it can be very inaccurate if that distribution is even slightly non-normal (Box 1953). Therefore, Bartlett's test is not recommended for routine use.

An approach that leads to tests that are much more robust to the underlying distribution is to transform the original values of the dependent variable to derive a *dispersion variable* and then to perform analysis of variance on this variable. The significance level for the test of homogeneity of variance is the *p*-value for the ANOVA *F*-test on the dispersion variable. Commonly used test for testing the homogeneity of variance using a dispersion variable is Levene Test given by Levene (1960). The procedure is described in the sequel.

Levene Test for homogeneity of Variances

The test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be. To conduct the Levene test, we divide the data into different groups based on the number of treatments if the error variance is either increasing or decreasing with the treatments, the residuals in the one treatment will tend to be more variable than those in others treatments. The Levene test than consists simply F – statistic based on one way ANOVA used to determine whether the mean of absolute/ Square root deviation from mean are significantly different or not. The residuals are obtained from the usual analysis of variance. The test statistic is given as

$$F = \frac{\left\{\sum_{i=1}^{v}(r_i - 1)\right\}}{v-1} \frac{\left\{\sum_{i=1}^{v} r_i(\bar{d}_{i.} - \bar{d}_{..})^2\right\}}{\sum_{i=1}^{v}\sum_{j=1}^{r_i}(d_{ij} - \bar{d}_{i.})^2} \sim F((v-1), \sum_{i=1}^{v}(r_i - 1))$$

where $d_{ij} = \left|e_{ij} - \bar{e}_i\right|$; $\bar{d}_{i.} = \frac{\sum_{j=1}^{r_i} d_{ij}}{r_i}$; $\bar{d}_{..} = \frac{\sum_{i=1}^{v}\sum_{j=1}^{r_i} d_{ij}}{\sum_{j=1}^{r_i} r_i}$ and e_{ij} is the j^{th} residual for the i^{th} plot, $\bar{e}_i$ is the mean of the residuals of the i^{th} treatment.

This test was modified by Brown and Forsythe (1974). In the modified test, the absolute deviation is taken from the median instead of mean in order to make the test more robust.

In the present investigation, the Bartlett's χ^2-test has been used for testing the homogeneity of error variances when the distribution of errors is normal and Levene test for non-normal errors.

Remark 1: In a block design**,** it can easily be shown that the sum of residuals within a given block is zero. Therefore, the residuals in a block of size 2 will be same with their sign reverse in order. As a consequence, *q* in Bartlett’s test and numerator in Levene test statistic becomes zero for the data generated from experiments conducted to compare only two treatments in a RCB design. Hence, the tests for homogeneity of error variances cannot be used for the experiments conducted to compare only two treatments in a RCB design. Inferences from such experiments may be drawn using Fisher-Behren t-test. Further, Bartlett’s test cannot be used for the experimental situations where some of the treatments are singly replicated.

Remark 2: In a RCB design, it can easily be shown that the sum of residuals from a particular treatment is zero. As a consequence, the denominator of Levene test statistic is zero for the data generated from RCB designs with two replications. Therefore, Levene test cannot be used for testing the homogeneity of error variances for the data generated from RCB designs with two replications.

Data Transformation

In this section, we shall discuss the remedial measures for non-normal and/or heterogeneous data in greater details.

Data transformation is the most appropriate remedial measure, in the situation where the variances are heterogeneous and are some functions of means. With this technique, the original data are converted to a new scale resulting into a new data set that is expected to satisfy the homogeneity of variances. Because a common transformation scale is applied to all observations, the comparative values between treatments are not altered and comparison between them remains valid.

Error partitioning is the remedial measure of heterogeneity that usually occurs in experiments, where, due to the nature of treatments tested some treatments have errors that are substantially higher (lower) than others.

Here, we shall concentrate on those situations where character under study is non-normal and variances are heterogeneous. Depending upon the functional relationship between variances and means, suitable transformation is adopted. The transformed variate should satisfy the following:

1. The variances of the transformed variate should be unaffected by changes in the means. This is also called the variance stabilizing transformation.
2. It should be normally distributed.
3. It should be one for which effects are linear and additive.
4. The transformed scale should be such for which an arithmetic average from the sample is an efficient estimate of true mean.

The following are the three transformations, which are being used most commonly, in biological research.

a) Logarithmic Transformation
b) Square root Transformation
c) Arc Sine or Angular Transformation

a) Logarithmic Transformation

This transformation is suitable for the data where the variance is proportional to square of the mean or the coefficient of variation (S.D./mean) is constant or where effects are multiplicative. These conditions are generally found in the data that are whole numbers and cover a wide range of values. This is usually the case when analyzing growth measurements such as trunk girth, length of extension growth, weight of tree or number of insects per plot, number of eggmass per plant or per unit area etc.

For such situations, it is appropriate to analyze *log X* instead of actual data, *X*. When data set involves small values or zeros, *log (X+1),* $log(2X+1)$ or $log\left(X+\frac{3}{8}\right)$ should be used instead of *log X*. This transformation would make errors normal, when observations follow negative binomial distribution like in the case of insect counts.

b) Square-Root Transformation

This transformation is appropriate for the data sets where the variance is proportional to the mean. Here, the data consists of small whole numbers, for example, data obtained in counting rare events, such as the number of infested plants in a plot, the number of insects caught in traps, number of weeds per plot, parthenocarpy in some varieties of mango. This data set generally follows the Poisson distribution and square root transformation approximates Poisson to normal distribution.

For these situations, it is better to analyze $\sqrt{X}$ than that of *X*, the actual data. If X is confirmed to small whole numbers then, $\sqrt{X+\frac{1}{2}}$ or $\sqrt{X+\frac{3}{8}}$ should be used instead of $\sqrt{X}$.

This transformation is also appropriate for the percentage data, where, the range is between 0 to 30% or between 70 and 100%.

c) Arc Sine Transformation

This transformation is appropriate for the data on proportions, *i.e.,* data obtained from a count and the data expressed as decimal fractions and percentages. The distribution of percentages is binomial and this transformation makes the distribution normal. Since the role of this transformation is not properly understood, there is a tendency to transform any percentage using arc sine transformation. But only that percentage data that are derived from count data, such as % barren tillers (which is derived from the ratio of the number of non-bearing tillers to the total number of tillers) should be transformed and not the percentage data such as % protein or % carbohydrates, %nitrogen, etc. which are not derived from count data. For these situations, it is better to analyze $\sin^{-1}(\sqrt{X})$than that of *X*, the actual data. If the value of X is 0%, it should be substituted by $\left(\frac{1}{4n}\right)$ and the value of 100% by $\left(100-\frac{1}{4n}\right)$, where *n* is the number of units upon which the percentage data is based.

It is interesting to note here that not all percentage data need to be transformed and even if they do, arc sine transformation is not the only transformation possible. The following rules may be useful in choosing the proper transformation scale for percentage data derived from count data.

Rule 1: The percentage data lying within the range 30 to 70% is homogeneous and no transformation is needed.

Rule 2: For percentage data lying within the range of either 0 to 30% or 70 to 100%, but not both, the square root transformation should be used.

Rule 3: For percentage that do not follow the ranges specified in Rule 1 or Rule 2, the Arc Sine transformation should be used.

The other transformations used are reciprocal square root [$\frac{1}{\sqrt{X}}$, when variance is proportional to cube of mean], reciprocal [$\frac{1}{X}$, when variance is proportional to fourth power of mean] and tangent hyperbolic transformation.

The transformation discussed above are a particular case of the general family of transformations known as Box-Cox transformation.

d) Box-Cox Transformation

By now we know that if the relation between the variance of observations and the mean is known then this information can be utilized in selecting the form of the transformation. We now elaborate on this point and show how it is possible to estimate the form of the required transformation from the data. The transformation suggested by Box and Cox (1964) is a power transformation of the original data. Let y_{ut} be the observation pertaining to the u^{th} plot; then the power transformation implies that we use y_{ut}*'s* as

$$y_{ut}^{*} = y_{ut}^{\lambda}.$$

The transformation parameter λ in $y_{ut}^{*} = y_{ut}^{\lambda}$ may be estimated simultaneously with the other model parameters (overall mean and treatment effects) using the method of maximum likelihood. The procedure consists of performing, for the various values of λ, a standard analysis of variance on

$$y_{ut}^{(\lambda)} = \begin{cases} \dfrac{y_{ut}^{\lambda} - 1}{\lambda \dot{y}_{ut}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}_{ut} \ln y_{ut} & \lambda = 0 \end{cases} \qquad \text{(A)}$$

where $\dot{y}_{ut} = \ln^{-1}\left[(1/n)\sum_{u=1}^{N}\sum_{t=1}^{n_u} \ln y_{ut}\right]$.

$\dot{y}_{ut}$ is the geometric mean of the observations. The maximum likelihood estimate of λ is the value for which the error sum of squares, say SSE (λ), is minimum. Notice that we cannot select the value of λ by directly comparing the error sum of squares from analysis of variance on y^{λ} because for each value of λ the error sum of squares is measured on a different scale. Equation (A) rescales the responses so that the error sums of squares are directly comparable. This is a very general transformation and the commonly used transformations follow as particular cases. The particular cases for different values of λ are given below.

λ	Transformation
1	No Transformation
½	Square Root
0	Log
-1/2	Reciprocal Square Root
-1	Reciprocal

Remark 3: If any one of the observations is zero then the geometric mean is undefined. In the expression (A), geometric mean is in denominator so it is not possible to compute that expression. For solving this problem, we add a small quantity to each of the observations.

Note: It should be emphasized that transformation, if needed, must take place right at the beginning of the analysis, all fitting of missing plot values, all adjustments by covariance etc. being done with the transformed variate and not with the original data. At the end, when the conclusions have been reached, it is permissible to 're-transform' the results so as to present them in the original units of measurement, but this is done only to render them more intelligible.

As a result of this transformation followed by back transformation, the means will rather be different from those that would have been obtained from the original data. A simple example is that without transformation, the mean of the numbers 1, 4, 9, 16 and 25 is 11. Suppose a square root transformation is used to give 1, 2, 3, 4 and 5, the mean is now 3, which after back-transformation gives 9. Usually the difference will not be so great because data do not usually vary as much as those given, but logarithmic and square root transformation always lead to a reduction of the mean, just as angles of equal formation usually lead to its moving away from the central value of 50%.

However, in practice, computing treatment means from original data is more frequently used because of its simplicity, but this may change the order of ranking of converted means for comparison. Although transformations make possible a valid analysis, they can be very awkward. For example, although a significant difference can be worked out in the usual way for means of the transformed data, none can be worked out for the treatment means after back transformation.

Non-parametric tests in the Analysis of Experimental Data

When the data remains non-normal and/or heterogeneous even after transformation, a recourse is made to non-parametric test procedures. A lot of attention is being paid to develop non-parametric tests for analysis of experimental data. Most of these non-parametric test procedures are based on rank statistic. The rank statistic has been used in development of these tests as the statistic based on ranks is

1. distribution free
2. easy to calculate and
3. simple to explain and understand.

Another reason for use of rank statistic is due to the well known result that the average rank approaches normality quickly as *n* (number of observations) increases, under the rather general conditions, while the same might not be true for the original data {see e.g. Conover and Iman (1976, 1981)}. The non-parametric test procedures available in literature cover completely randomized designs, randomized complete block designs, balanced incomplete block designs, design for bioassays, split plot designs, cross-over designs and so on. For an excellent and elaborate discussions on non-parametric tests in the analysis of experimental data, one may refer to Siegel and Castellan Jr. (1988), Deshpande, Gore and Shanubhogue (1995), Sen (1996), and Hollander and Wolfe (1999).

Kruskal-Wallis Test can be used for the analysis of data from completely randomized designs. Skillings and Mack Test helps in analyzing the data from a general block design. Friedman Test and Durbin Test are particular cases of this test. Friedman Test is used for the analysis of data from randomized complete block designs and Durbin test for the analysis of data from balanced incomplete block designs.

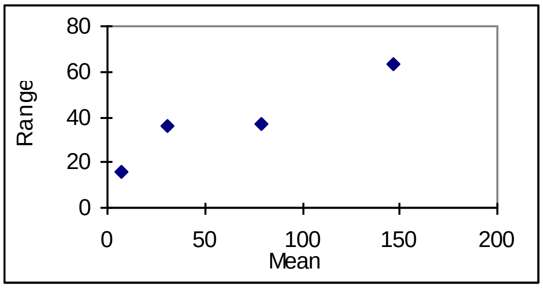
Some examples of testing the assumptions of normality and homogeneity of errors and remedial measures are discussed in the Appendix.

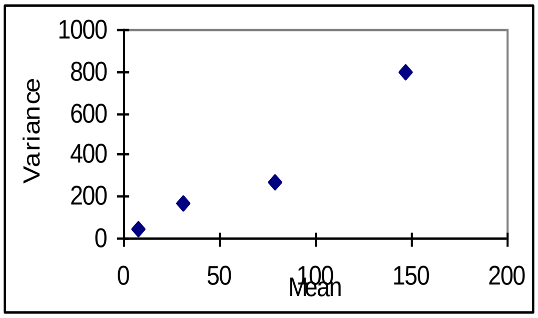
Appendix

Example 1: Suppose an entomologist is interested in determining whether four different kinds of traps caught equivalent insects when applied to same field. Each of the traps is used six times on the field and resulting data (number of insects per hour) are as shown below alongwith mean, variance and range.

Treatment	Replication						Mean	Variance	Range
	I	II	III	IV	V	VI	$\bar{Y}_i$	S_i^2	
A	3	1	12	7	17	2	7	40.4	16
B	9	29	21	24	28	45	31	138.4	36
C	63	84	97	61	98	71	79	270.8	37
D	172	118	109	172	143	168	147	798.4	63

A scatter plot of mean and variance and mean versus range are given as follows:





Both plots indicate that variances are heterogeneous and variance is proportional to mean.

Obtain the residuals for testing the normality and homogeneity of error terms. The residuals obtained are given below:

Treatment	Replication						Mean	Variance
	I	II	III	IV	V	VI		S_i^2
A	-1.00	0.75	10.00	-1.25	3.25	-11.75	0	50.35
B	-14.00	9.75	0.00	-3.25	-4.75	12.25	0	94.85
C	-13.00	11.75	23.00	-19.25	12.25	-14.75	0	314.85
D	28.00	-22.25	-33.00	23.75	-10.75	14.25	0	650.20

Test for Normality of error terms

Shapiro-Wilk Test		Kolmogrov-Smirnov Test	
Statistic (SW)	**p-value**	**Statistic (KS)**	**p-value**
0.980	0.882	0.110	0.200

The errors were found to be normally distributed. Therefore, homogeneity of error variances was tested using Bartlett's test. It is described in the sequel.

Pooled Variance $\left(S_p^2\right) = \frac{5(50.35+94.85+314.85+650.20)}{20} = 277.5625$

$$q = 20\log_{10} 277.5625 - 5\left[\log_{10} 50.35 + \log_{10} 64.85 + \log_{10} 314.85 + \log_{10} 650.20\right]$$
$$= 3.916278$$
$$c = 1 + \frac{1}{9}\left(\frac{4}{5} - \frac{1}{20}\right) = 1.08333$$
$$\chi_0^2 = 8.324.$$

Since $\chi_{0.05,3}^2 = 7.81$, therefore, we reject the null hypothesis and conclude that the variances are unequal.

The $\frac{S_i^2}{\bar{Y}_{i.}}$ are 5.77, 5.32, 3.43 and 5.43, indicating that variance is proportional to mean. Therefore, square root transformation should be used. After application of square root transformation, the residuals are

Treatment	Replication						Variance
	I	II	III	IV	V	VI	S_i^2
A	-0.03614	-0.92542	1.05800	0.20614	0.98287	-1.28544	0.928
B	-1.34939	0.87854	-0.40473	-0.12183	-0.42993	1.42735	0.999
C	-0.28226	0.78841	0.99143	-1.08068	0.30794	-0.72483	0.694
D	1.66779	-0.74153	-1.64469	0.99637	-0.86087	0.58293	1.622

Normality of error terms on the transformed data:

Shapiro-Wilk Test		Kolmogrov-Smirnov Test	
Statistic (SW)	**p-value**	**Statistic (KS)**	**p-value**
0.956	0.414	0.127	0.200

The errors remain normally distributed after transformation. The results of homogeneity of error variances using Bartlett's test are

Bartlett's Test (normal distribution): Test statistic = 0.89, p-value = 0.828
Hence, we conclude that the errors are normally distributed and have a constant variance after transformation.

The results of analysis of variance with original and transformed data are given in the sequel.

ANOVA: Original Data

Source	DF	Seq SS	Adj. SS	Mean Square	F (F-calc)	p(Pr>F)
Replication	5	689.0	689.0	137.8	0.37	0.86
Treatment	3	70828.5	70828.5	23609.5	63.80	0.00
Error	15	5551.0	5551.0	370.1		
Total	23	77068.5				

R-Square	Root MSE
92.80%	19.2371

Tukey Simultaneous Tests for All Pairwise Treatment Comparisons

	1	2	3	4
1	.			
2	0.3525	.		
3	0.0001	0.0013	.	
4	0.0000	0.0000	0.0001	.

ANOVA: Transformed Data

Source	DF	Seq SS	Adj. SS	Mean Square	F (F-calc)	p(Pr>F)
Replication	5	5.055	5.055	1.011	0.71	0.622
Treatment	3	326.603	326.603	108.868	76.98	0.000
Error	15	21.214	21.214	1.414		
Total	23	352.872				

R-Square	Root MSE
93.99%	1.18922

Tukey Simultaneous Tests for All Pairwise Treatment Comparisons

	1	2	3	4
1	.			
2	0.0091			
3	0.0000	0.0003		
4	0.0000	0.0000	0.0015	.

With transformed data treatments 1 and 2 are significantly different whereas with original data, they were not.

Example 2: A varietal trial on Rapeseed-Mustard was conducted at Faizabad with 11 varieties using a randomized complete block design with 3 replications. The experimental data (Yield in kg/ha) obtained from the above experiment is

Treatments ↓	**Replications→**		
	R1	**R2**	**R3**
MCN-157	952.380	1058.200	1079.364
MCN-158	846.560	634.920	687.830
MCN-159	529.100	687.830	687.830
MCN-160	1058.200	1005.290	952.380
MCN-161	1111.110	888.888	846.560

MCN-162	899.470	634.920	1005.290
MCN-163	1058.200	1164.020	952.380
MCN-164	687.830	740.740	529.100
MCN-165	952.380	952.380	867.724
MCN-166	1058.200	1058.200	529.100
MCN-167	1269.840	1164.020	1216.930

The analysis of variance of the original data is given as

ANOVA: Original Data

Sources	DF	SS	MS	F	Prob. >F
Replication	2	52534.9880	26267.4940	1.46	0.2563
Treatment	10	967055.0471	96705.5047	5.37	0.0007
Error	20	360218.589	18010.929		
Total	**32**	1379808.624			

R-Square	CV	RMSE	Yld Mean
0.738936	14.878	134.2048	902.035

Normality of error terms was tested, the results are given as

Shapiro-Wilk Test		Kolmogrov-Smirnov Test	
Statistic (SW)	p-value	Statistic (KS)	p-value
0.9679	0.4249	0.1018	>0.1500

Since the data is normal, therefore, Bartlett's test is used for testing the homogeneity of error variances. The results are given as

Bartlett's Test
Test Statistic : 20.177
P-Value : 0.0276

The errors were found to be heterogeneous.

Therefore, we can conclude that the data is heterogeneous and normal.

Therefore, Box-Cox transformation was used as a remedial measure. In the sequel we describe the results of the Box-Cox transformation.

For this we transform the data by varying λ from -10 to +10 with an increment of 0.01. The error sum of squares are computed for each value of λ. The value of λ with minimum error sum of squares is used for transformation given in (A). The minimum value SSE is obtained for λ = 2.38. Therefore, reciprocal transformation was used.

The assumptions of normality and homogeneity of errors are again tested using the transformed data.

Normality of error terms was tested, the results are given as

Shapiro-Wilk Test		Kolmogrov-Smirnov Test	
Statistic (SW)	p-value	Statistic (KS)	p-value
0.984	0.8885	0.0867	>0.1500

Since the data is normal, therefore, Bartlett’s Test is used for testing the homogeneity of error variances. The results are given as

Bartlett's Test (normal distribution)	
Test Statistic	: 15.725
P-Value	: 0.107757

The transformed observations were found to be normal and homogeneous Therefore, ANOVA was performed on the transformed data. The results obtained are:

ANOVA: Transformed Data

Sources	DF	SS	MS	F	Prob. >F
Replication	2	3.865471E13	1.93273335E13	1.62	0.2238
Treatment	10	7.8841391E14	7.8841391E13	6.59	0.0002
Error	20	2.3934391E14	1.1967195E13		
Total	**32**	**1.0664125E15**			

R-Square	CV	RMSE	Transformed Yld Mean
0.7756	29.563	3459363	11701777

We can see that there is no change in the results of significance of treatment and replication effects. However, the transformed data satisfied the assumptions of ANOVA.

NONPARAMETRIC TESTS

1. Introduction

A parametric test specifies certain conditions about the distribution of responses in the population from which the research sample was drawn. The meaningfulness of the results of a parametric test depends on the validity of these assumptions. A nonparametric test is based on a model that specifies very general conditions and none regarding the specific form of the distribution from which the sample was drawn. Hence nonparametric tests are also known as distribution free tests. Certain assumptions are associated with most nonparametric statistical tests, but these are fewer and weaker than those of parametric tests.

Nonparametric test statistics utilize some simple aspects of sample data such as the signs of measurements, order relationships or category frequencies. Therefore, stretching or compressing the scale does not alter them. As a consequence, the null distribution of the nonparametric test statistic can be determined without regard to the shape of the parent population distribution. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and, as a result, if the data have one or two outliers, their influence is negated.

Advantages of nonparametric tests

- Non-parametric methods are used with all scales
- When sample size is very small, there may be no alternative to use a nonparametric test unless the population distribution is known exactly
- They are easier to learn and compute
- Fewer assumptions are made
- Due to the reliance on fewer assumptions, non-parametric methods are more robust
- Need not involve population parameters
- Results may be as exact as parametric procedures

Disadvantages of nonparametric tests

- There may be wastage of information
- Parametric models are more efficient if data permit.
- It is difficult to compute by hand for large samples
- Tables are not widely available
- In cases where a parametric test would be appropriate, non-parametric tests have less power. In other words, a larger sample size can be required to draw conclusions with the same degree of confidence.

The inferences drawn from tests based on the parametric tests such as t, F and χ^2 may be seriously affected when the parent population’s distribution is not normal. The adverse effect could be more when sample size is small. Thus when there is doubt about the distribution of the parent population, a nonparametric method should be used. In many situations, particularly in social and behavioral sciences, observations are difficult or impossible to take on numerical scales and a suitable nonparametric test is an alternative under such situations. Some commonly used nonparametric tests are discussed in the sequel.

2. Run Test for Randomness

Run test is used for examining whether or not a set of observations constitutes a random sample from an infinite population. Test for randomness is of major importance because the assumption of randomness underlies statistical inference. In addition, tests for randomness are important for time series analysis. Departure from randomness can take many forms.

H_0: Sample values come from a random sequence
H_1: Sample values come from a non-random sequence

Test Statistic: Let r be the number of runs (a run is a sequence of signs of same kind bounded by signs of other kind). For finding the number of runs, the observations are listed in their order of occurrence. Each observation is denoted by a '+' sign if it is more than the previous observation and by a '-' sign if it is less than the previous observation. Total number of runs up (+) and down (-) is counted. Too few runs indicate that the sequence is not random (has persistency) and too many runs also indicate that the sequence is not random (is zigzag).

Critical Value: Critical value for the test is obtained from the table for a given value of n and at desired level of significance (α). Let this value be r_c.

Decision Rule: If r_c (lower) $\leq r \leq$ r_c (upper), accept H_0. Otherwise reject H_0.

Tied Values: If an observation is equal to its preceding observation denote it by zero. While counting the number of runs ignore it and reduce the value of n accordingly.

Large Sample Sizes: When sample size is greater than 25 the critical value r_c can be obtained using a normal distribution approximation.

The critical values for two-sided test at 5% level of significance are

r_c (lower) $= \mu - 1.96\ \sigma$
r_c (upper) $= \mu + 1.96\ \sigma$

For one-sided tests, these are

r_c (left tailed) $= \mu - 1.65\ \sigma$, if $r \leq r_c$, reject H_0
r_c (right tailed) $= \mu + 1.65\ \sigma$, if $r \geq r_c$, reject H_0,

where $\mu = \frac{2n-1}{3}$ and $\sigma = \sqrt{\frac{16n-29}{90}}$.

Example 2.1: Data on value of imports of selected agricultural production inputs from U.K. by a county (in million dollars) during recent 12 years is given below: Is the sequence random?

5.2	5.5	3.8	2.5	8.3	2.1	1.7	10.0	10.0	6.9	7.5	10.6

H_0: Sequence is random.
H_1: Sequence is not random.

5.2	5.5	3.8	2.5	8.3	2.1	1.7	10.0	10.0	6.9	7.5	10.6
	+	-	-	+	-	-	+	0	-	+	+

Here n = 11, the number of runs r = 7. Critical n values for α = 5% (two sided test) from the table are r_c (lower) = 4 and r_c (upper) = 10. Since r_c (lower) $\leq$ r $\leq$ r_c (upper), i.e., observed r lies between 4 and 10, H_0 is accepted. The sequence is random.

3. Wald-Wolfowitz Two-Sample Run Test

Wald–Wolfowitz run test is used to examine whether two random samples come from populations having same distribution. This test can detect differences in averages or spread or any other important aspect between the two populations. This test is efficient when each sample size is moderately large (greater than or equal to 10).

H_0: Two sample come from populations having same distribution
H_1: Two sample come from populations having different distributions

Test Statistic: Let r denote the number of runs. To obtain r, list the $n_1 + n_2$ observations from two samples in order of magnitude. Denote observations from one sample by x's and other by y's. Count the number of runs.

Critical Value: Difference in location results in few runs and difference in spread also result in few number of runs. Consequently, critical region for this test is always one-sided. The critical value to decide whether or not the number of runs are few, is obtained from the table. The table gives critical value r_c for n_1 (size of sample 1) and n_2 (size of sample 2) at 5% level of significance.

Decision Rule: If r $\leq$ r_c, reject H_0.

Tie: In case *x* and *y* observations have same value, place the observation x(y) first if run of x(y) observation is continuing.

Large Sample Sizes: For sample sizes larger than 20 critical value r_c is given below.

$r_c = \mu - 1.96\,\sigma$ at 5% level of significance

where $\mu = 1 + \dfrac{2n_1n_2}{n_1 + n_2}$ and $\sigma = \sqrt{\dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$

Example 3.1: To determine if a new hybrid seeding produces a bushier flowering plant, following data was collected. Examine if the data indicate that new hybrid produces larger shrubs than the current variety?

Shrubs Girth (in inches)								
Hybrid	x	31.8	32.8	39.2	36.0	30.0	34.5	37.4
Current Variety	y	35.5	27.6	21.3	24.8	36.7	30.0	

H_0: x and y populations are identical
H_1: There is some difference in girth of x and y shrubs.

Consider the combined ordered data.

21.3	24.8	27.6	30.0	30.0	31.8	32.8	34.5	35.5	36.0	36.7	37.4	39.2
y	y	y	y	x	x	x	x	y	x	y	x	x

Test statistic r = 6 (total number of runs). For $n_1 = 7$ and $n_2 = 6$, critical value r_c at 5% level of significance is 3. Since $r > r_c$, we accept H_0 and conclude that x and y have identical distribution.

4. Median Test for Two Samples

To test whether or not two samples come from same population, median test is used. It is more efficient than the run test but each sample should be of size 10 at least. In this case, the hypothesis to be tested is

H_0 : Two samples come from populations having same distribution.
H_1 : Two samples come from populations having different distribution.

Test Statistic: χ^2 (Chi-square). To test the value of test statistics two samples of sizes n_1 and n_2 are combined. Median M of the combined sample of size $n = n_1 + n_2$ is obtained. Number of observations below and above the median M for each sample is determined. This is then analyzed as a 2 × 2 contingency table in the manner given below.

	Number of Observations		
	Sample 1	Sample 2	Total
Above Median	a	b	a+b
Below Median	c	d	c+d
	$a+c= n_1$	$b+d = n_2$	n = a+b+c+d

$$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + c)(b + d)(a + b)(c + d)}$$

Decision Rule: if $\chi^2 \geq \chi^2_c$, reject H_0 otherwise accept it.

Tie: Ties are ignored and n is adjusted accordingly.

Remark: This test can be extended to *k* samples with number of observations below and above the combined median M from a 2 × k contingency table.

Example 4.1: Perform a median test on the problem of Example 3.1 for testing that the two samples come from same population.

H_0 : x and y populations are identical.
H_1 : There is some difference in girth of x and y shrubs.

Seventh value 32.8 is the median of combined ordered sequence.

	Number of Observations		
	x	y	Total
Above M	4	2	6
Below M	2	4	6
	6	6	12

$$\chi^2 = \frac{12(16-4)^2}{6.6.6.6} = \frac{4}{3} = 1.33.$$

Since $\chi^2 = 1.33 < \chi_c^2 = 3.84$, H_0 is accepted. It is concluded that two samples come from the same population. There is no significant difference in the girth of hybrid and current variety of shrub.

Remark: This example is for demonstrating the test procedure. In real situation n should be at least 20 and each cell frequency at least 5.

5. Sign Test for Matched Pairs

In many situations, comparison of effect of two treatments is of interest but observations occur in pairs. Thus the two samples are not truly random. Because of such pair-wise dependence ordinary two sample tests are not appropriate. In such situations when one member of the pair is associated with the treatment A and the other with treatment B, nonparametric sign test has wide applicability. It can be applied even when qualitative data are available. As the name suggests it is based on the signs of the response differences D_i. If i^{th} pair of observation is denoted by (x_i, y_i) where x is the effect of treatment A and y to B then $D_i = x_i - y_i$. The hypothesis to be tested is

H_0 : No difference in the effect of treatments A and B.
H_1 : A is better than B.

Test Statistic: Let S be the number of '-' signs.

Critical Value: Critical value S_c corresponding to n, the number of pairs, is given in Table 3. Significance level is given by α_1 as critical region is one sided (left tailed).

Decision Rule: If $S \le S_c$ reject H_0, otherwise accept H_0.

Tie: In case two values of a pair are equal, reject that pair and reduce the number of observations accordingly.

Remark: In case, if the alternative H_1 is that there is some difference in effect of A and B, S represents either the number of negative signs or the number of positive signs whichever turn out to be smaller. Critical region is two sided and significance level is given by α_2 for finding S_c.

Example 5.1: In a market study, two brands of lemonade were compared. Each of 50 judges tasted two samples, one of brand A and one of brand B with the following results: 35 preferred brand A, 10 preferred B, and 5 could not tell the difference. Thus, n = 45 and S = 10. Assuming α_1 = 5%, critical value S_c = 16 from Table 3. Since $S < S_c$, we reject H_0 of no difference in favour of the alternative H_1 that the brand A is preferred.

6. Wilcoxon Signed Rank Test for Matched Pairs

In situations where there is some kind of pairing between observations in the two samples, ordinary two sample tests are not appropriate. Signed rank tests are useful in such situations. When observations are measured data, signed rank test is more efficient than sign test as it takes account of the magnitude of the observed differences, if the difference between the response of the two treatments A and B is to be tested the test hypothesis is

H_0 : No difference in the effect of treatments A and B.
H_1 : Treatment A is better than B.

Test Statistic: T represents the sum of ranks with negative signs. For calculating T, obtain the differences $D_i = x_i - y_i$ where x_i's are response of treatment A and y_i's of treatment B. Rank the absolute values of differences. Smallest give rank 1. Ties are assigned average ranks. Assign to each rank sign of observed difference. Obtain the sum of negative ranks.

Critical Value: T_c is given in Table 4 for n number of pairs. Significance level is given by α_1 as critical region is one sided.

Decision Rule: $T \leq T_c$ reject H_0, other wise accept it.

Tie: Discard the pair for which difference = 0 and reduce n accordingly. Equal differences are assigned average ranks.

Example 6.1: Blood pressure reading of ten patients before and after medication for reducing the blood pressure are as follows:

Patient		1	2	3	4	5	6	7	8	9	10
Before treatment	x	86	84	78	90	92	77	89	90	90	86
After treatment	y	80	80	92	79	92	82	88	89	92	83
Differences		6	4	-14	11	0	-5	1	1	-2	3
Rank		7	5	9	8	Discard	6	1.5	1.5	3	4
Sign		+	+	-	+	Discard	-	+	+	-	+

Test the null hypothesis of no effect against the alternative that medication is effective.

Rank sum of negative differences = 3+6+9 = 18. Therefore value of test statistic T = 18. For n = 9 and α_1 = 5%, T_c = 8 from Table 4. Since $T > T_c$, null hypothesis of no effect of medication is accepted.

7. Kolmogorov-Smirnov Test

In situations where there is unequal number of observations in two samples, Kolmogorov-Smirnov test is appropriate. This test is used to test whether there is any significant difference between two treatments A and B (say). The test hypothesis is

H_0 : No difference in the effect of treatments A and B.
H_1 : There is some difference in the effect of treatments A and B.

Test Statistic: The test statistic is $D_{m,n} = \sup|F_m(x) - G_n(x)|$, F and G are the sample empirical distributions of sample observations of two samples respectively with respective sample sizes *m* and *n*. $F(x_i)$ is calculated as the average number of sample observations of the first sample that are less than x_i. Similarly $G(x_i)$ is calculated. $D_{m,n}$ is largest value of the absolute difference between F(x) and G(x).

Critical Value: Tabulated value of $D_{m,n}$ is available for different values of m, n and for different levels of significance and is given in Table 4 for n number of pairs. Significance level is given by α_1 as critical region is one-sided.

*Decision Rule***:** If the calculated value of $D_{m,n}$ is greater than the tabulated value of $D_{m,n}$, H_0 is rejected otherwise it is accepted.

Example 7.1: The following data represent the lifetimes (hours) of batteries for different brands:

Brand A	40	30	40	45	55	30
Brand B	50	50	45	55	60	40

Are these brands different with respect to average life?

We first calculate the sample empirical distributions of two samples as follows:

x	$F_6(x)$	$G_6(x)$	$\lvert F_6(x)-G_6(x)\rvert$
30	2/6	0	2/6
40	4/6	1/6	3/6
45	5/6	2/6	3/6
50	5/6	4/6	1/6
55	1	5/6	1/6
60	1	1	0

$D_{6,6} = \sup\lvert F_6(x)-G_6(x)\rvert = 3/6$. From table, the critical value for m = n = 6 at level $\alpha = .05$ is 4/6. Since the calculated value of $D_{m,n}$ is not greater than the tabulated value, H_0 is not rejected and it is concluded that the average length of life for two brands is the same.

8. Kruskal-Wallis Test

This test is appropriate for use under the following circumstances: (a) If somebody wants to compare three or more conditions; (b) each condition is performed by a *different* group of participants; i.e. you have an independent-measures design with three or more conditions. (c) data do not meet the requirements for a parametric test. (i.e. use it if the data are not normally distributed; if the variances for the different conditions are markedly
different; or if the data are measurements on an ordinal scale).

If the data meet the requirements for a parametric test, it is better to use a one-way independent-measures Analysis of Variance (ANOVA) because it is more powerful than the Kruskal-Wallis test.

Example: Does physical exercise alleviate depression? Here, some individuals are randomly allocated to one of three groups: no exercise; 20 minutes of jogging per day; or 60 minutes of jogging per day. At the end of a month, ach individual is asked to rate how depressed they now feel, on a *Likert scale* that runs from 1 ("totally miserable") through to 100 (ecstatically happy").

Rating on depression scale

No exercise	Jogging for 20 minutes	Jogging for 60 minutes
23	22	59
26	27	66
51	39	38

49	29	49
58	46	56
37	48	60
29	49	56
44	65	62

Out Put

Test Statistics[a,b]

	Depression
Chi-Square	7.290
df	2
Asymp. Sig.	.026

a. Kruskal Wallis Test

b. Grouping Variable: Depression

A Kruskal-Wallis test revealed that there is a significant effect of exercise on depression.

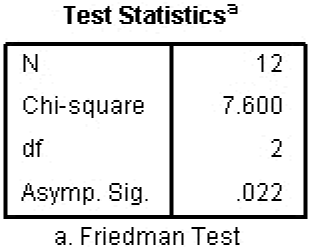
9. Friedman’s Test

It is a non-parametric statistical test for testing whether samples originate from the same distribution. It is used for comparing more than two samples that are related. When the Friedman’s test leads to significant results, then at least one of the samples is different from the other samples.

Example: A researcher wants to examine whether music has an effect on the perceived psychological effort required to perform an exercise session. The dependent variable is "perceived effort to perform exercise" and the independent variable is "music type", which consists of three categories: "no music", "classical music" and "dance music". To test whether music has an effect on the perceived psychological effort required to perform an exercise session, the researcher recruited 12 runners who each ran three times on a treadmill for 30 minutes. For consistency, the treadmill speed was the same for all three runs. In a random order, each subject ran: (a) listening to no music at all; (b) listening to classical music; and (c) listening to dance music. At the end of each run, subjects were asked to record how hard the running session felt on a scale of 1 to 10, with 1 being easy and 10 extremely hard. A Friedman test was then carried out to see if there were differences in perceived effort based on music type.

No Music	Classical Music	Dance Music
8	8	7
7	6	6
6	8	6
8	9	7
5	8	5
9	7	7
7	7	7
8	7	7

8	6	8
7	6	6
7	8	6
9	9	6



Test Statistics[a]

N	12
Chi-square	7.600
df	2
Asymp. Sig.	.022

a. Friedman Test

It shows that an overall statistically significant difference between the mean ranks of the related groups.

Table 1: Critical values for runs up and down test

	$\alpha_1 = 5$ %		$\alpha_1 = 2.5$ %		$\alpha_1 = 1$ %		$\alpha_1 = 0.5$ %	
	$\alpha_2 = 10$ %		$\alpha_2 = 5$ %		$\alpha_2 = 2$ %		$\alpha_2 = 1$ %	
n	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
3	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-
5	1	-	1	-	-	-	-	-
6	1	-	1	-	1	-	1	-
7	2	-	2	-	1	-	1	-
8	2	-	2	-	2	-	1	-
9	3	8	3	-	3	-	2	-
10	3	9	3	-	3	-	2	-
11	4	10	4	10	3	-	3	-
12	4	11	4	11	4	-	3	-
13	5	12	5	12	4	12	4	-
14	6	12	5	13	5	13	4	13
15	6	13	6	14	5	14	4	14
16	7	14	6	14	6	15	5	15
17	7	15	7	15	6	16	6	16
18	8	15	7	16	7	16	6	17
19	8	16	8	17	7	17	7	18
20	9	17	8	17	8	18	7	18
21	10	18	9	18	8	19	8	19
22	10	18	10	19	9	20	8	20
23	1	19	10	20	10	20	9	21
24	1	20	11	20	10	21	10	22
25	12	21	11	21	11	22	10	22

α_1 : Significance level for one sided test
α_2 : Significance level for two sided test

*Source***:** Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

Table 2: Critical values for the two sample run test.

n_1 n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	2	2	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

Significance level 5%

Source: Statistics in Research by Borten Ostle. Ames. Iowa USA. Iowa State University Press.

Table 3: Critical values for the Sign test (matched pairs)

	α_1	5 %	2.5 %	1 %	0.5 %		α_1	5 %	2.5 %	1 %	0.5 %
n	α_2	10 %	5 %	2 %	1 %	n	α_2	10 %	5 %	2 %	1 %
1		-	-	-	-	26		8	7	6	6
2		-	-	-	-	27		8	7	7	6
3		-	-	-	-	28		9	8	7	6
4		-	-	-	-	29		9	8	7	7
5		0	-	-	-	30		10	9	8	7
6		0	0	-	-	31		10	9	8	7
7		0	0	0	-	32		10	9	8	8
8		1	0	0	0	33		11	10	9	8
9		1	1	0	0	34		11	10	9	9
10		1	1	0	0	35		12	11	10	9
11		2	1	1	0	36		12	11	10	9
12		2	2	1	1	37		13	12	10	10
13		3	2	1	1	38		13	12	11	10
14		3	2	2	1	39		13	12	11	11
15		3	3	2	2	40		14	13	12	11
16		4	3	2	2	41		14	13	12	11
17		4	4	3	2	42		15	14	13	12
18		5	4	3	3	43		15	14	13	12
19		5	4	4	3	44		16	15	13	13
20		5	5	4	3	45		16	15	14	13
21		6	5	4	4	46		16	15	14	13
22		6	5	5	4	47		17	16	15	14
23		7	6	5	4	48		17	16	15	14
24		7	6	5	5	49		18	17	15	15
25		7	7	6	5	50		18	17	16	15

α_1 : Significance level for one sided test
α_2 : Significance level for two sided test

*Source***:** Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

Table 4: Critical values for the Wilcoxon signed rank test

	α_1	5 %	2.5 %	1 %	0.5 %		α_1	5 %	2.5 %	1 %	0.5 %
n	α_2	10 %	5 %	2 %	*1 %*	*n*	α_2	10 %	5 %	2 %	1 %
1		-	-	-	-	26		110	98	84	75
2		-	-	-	-	27		119	107	92	83
3		-	-	-	-	28		130	116	101	91
4		-	-	-	-	29		140	126	110	100
5		0	-	-	-	30		151	137	120	109
6		2	0	-	-	31		163	147	130	118
7		3	2	0	-	32		175	159	140	128
8		5	3	1	0	33		187	170	151	138
9		8	5	3	1	34		200	182	162	148
10		10	8	5	3	35		213	195	173	159
11		13	10	7	5	36		227	208	185	171
12		17	13	9	7	37		241	221	198	182
13		21	17	12	9	38		256	235	211	194
14		25	21	15	12	39		271	239	224	207
15		30	25	19	15	40		286	264	238	220
16		35	29	23	19	41		302	279	252	233
17		41	34	27	23	42		319	294	266	244
18		47	40	32	27	43		336	310	281	261
19		53	46	37	32	44		353	327	296	276
20		60	52	43	37	45		371	343	312	291
21		67	58	49	42	46		389	361	328	307
22		75	65	55	48	47		407	278	345	322
23		83	73	62	54	48		426	296	362	339
24		91	81	69	61	49		446	415	379	355
25		100	89	76	68	50		466	434	397	373

α_1 : Significance level for one sided test
α_2 : Significance level for two sided test

Source: Distribution Free Tests by H.R. Neave and P.L. Worthington. London, Unwin Hyman.

SAMPLING IN FIELD EXPERIMENTS

In agricultural field experiments, the size of the plot is selected in order to achieve a prescribed degree of precision for measurement of the character of primary interest. We then measure the character under study on the whole of the experimental unit i.e. plot. Because of the nature of the character of primary interest like yield, the plot size required is often larger than that needed to measure other characters. In order to save expense and time the measurements of additional characters of interest can be made by sampling a fraction of the whole plot. For example, for plant height, the measurements can be made only from say 10 of the 200 plants in the plot, for tiller number, count only 1 m^2 of the 15 m^2 plot, for leaf area, measure from only 20 of the approximately 2000 leaves in the plot. For such cases like plant height, leaf area etc. it may not be always feasible or desirable to get the plot wise measurements. Here we resort to sampling in each plot and obtain the measurements on a certain number of sampling units in each plot and subject the data for statistical analysis.

An appropriate sample is one that provides an estimate, or a sample value, that is as close as possible to the value that would have been obtained had all plants in the plot been measured - the plot value. The difference between the sample value and the plot value constitutes the sampling error. Thus a good sampling technique is one that gives small sampling error.

The **sampling unit** is the unit on which actual measurement is made. The important features of an appropriate sampling unit are:

- **Ease of Identifications**
- **Ease of Measurement**
- **High Precision**
- **Low Cost**

The number of sampling units taken from the population is **sample size**. In a replicated field trial where each plot is a population, sample size could be the number of plants per plot used for measuring plant height, the number of leaves per plot used for measuring leaf area, or the number of hills per plot used for counting tillers. The required sample size for a particular experiment is governed by:

(i) The size of the variability among sampling units within the same plot (sampling variance)

(ii) The degree of precision desired for the character of interest.

In practice, the size of the sampling variance for most plant characters is generally not known. The desired level of precision can, however, be prescribed by the researcher based on experimental objective and previous experience, in terms of the margin of error, either of the plot mean or of the treatment mean.

The sample size for a simple random sampling design that can satisfy a prescribed margin of error of the plot mean is computed as:

$$n = \frac{(Z_{\alpha}^2)(v_s)}{(d^2)(\bar{X}^2)}$$

where n is the required sample size, Z_α is the value of the standardized normal variate corresponding to the level of significance α , v_s is the sampling variance, $\overline{X}$ is the mean value, and d is the margin of error expressed as a fraction of the plot mean.

The information of primary interest to the researcher is usually the treatment means (the average over all plots receiving the same treatment) or actually the difference of means, rather than the plot mean (the value from a single plot). Thus, the desired degree of precision is usually specified in terms of the margin of error of the treatment mean rather than of the plot mean. In such a case, sample size is computed as:

$$n = \frac{(Z_\alpha^2)(v_s)}{r(D^2)(\overline{X}^2) - (Z_\alpha^2)(v_p)}$$

where n is the required sample size, r is the number of replications, Z_α and v_s are as defined earlier, v_p is the variance between plots of the same treatment (i.e. experimental error), and D is the prescribed margin of error expressed as a fraction of the treatment mean. In this case, additional information on the size of the experimental error (v_p) is needed to compute sample size.

A **sampling design** specifies the manner in which the n sampling units are to be selected from the whole plot. There are five commonly used sampling designs in replicated field trials: simple random sampling, multistage random sampling, stratified random sampling, stratified multistage random sampling, and sub-sampling with an auxiliary variable.

In a **simple random sampling** design, there is only one type of sampling unit and, hence, the sample size (n) refers to the total number of sampling units to be selected from each plot consisting of N units. The selection of the n sampling units is done in such a way that each of the N units in the plot is given the same chance of being selected in plot sampling, two of the most commonly used random procedures for selecting n sampling units per plot are the random-number technique and the random - pair technique.

In contrast to the simple random sampling design, where only one type of sampling unit is involved, the **multistage random sampling** design is characterized by a series of sampling stages. Each stage has its own unique sampling unit. This design is suited for cases where the sampling unit is not the same as the measurement unit. For example, in a rice field experiment, the unit of measurement for panicle length is a panicle and that for leaf area is a leaf. The use of either the panicle or the leaf as the sampling unit, however, would require the counting and listing of all panicles or all leaves in the plot which is time-consuming task that would definitely not be practical.

The **stratified random sampling design** is useful where there is large variation between sampling units and where important sources of variability follow a consistent pattern. In such cases, the precision of the sample estimate can be improved by grouping the sampling units into different strata in such a way that variability between sampling units within a stratum is smaller than that between sampling units from different strata. Some examples of stratification criterion used in agricultural experiments are as follows:

- **Soil Fertility Pattern**. In an insecticide trial where block is based primarily on the direction of insect migration, known patterns of soil fertility cause substantial variability among plants in the same plot. In such a case, a stratified random sampling design may be used so that each plot is first divided into several strata based on the known fertility patterns and sample plants are then randomly selected from each stratum.

- **Stress Level**. In a variety screening trial for tolerance for soil salinity, areas within the same plot may be stratified according to the salinity level before sample plants are randomly selected from each stratum.
- **Within-Plant Variance**. In a rice hill, panicles from the taller tillers are generally larger than those from the shorter ones. Hence, in measuring such yield components as panicle length or number of grains per panicles, panicles within a hill are stratified according to the relative height of the tillers before sample panicles are randomly selected from each position (or stratum).

Stratified multistage random sampling: Consider the case where a rice researcher wishes to measure the average number of grains per panicle through the use of a two-stage sampling design with individual hills in the plot as the primary sampling unit and individual panicles in a hill as the secondary sampling unit. It is realized that the number of grains per panicle varies greatly between the different panicles of the same hill. A logical alternative is to apply the stratification technique by dividing the panicles in each selected hill (i.e., primary sampling unit) into k strata, based on their relative position in the hill, before a simple random sample of m panicles from each stratum is taken separately and independently for the k strata. In this case, the sampling technique is based on a **two-stage sampling design with stratification** applied on the secondary unit. Of course, instead of the secondary unit (panicles) the researcher could have stratified the primary unit (i.e., single-hill) based on any source of variation pertinent to his experiment. In that case, the sampling technique would have been a two-stage sampling design with stratification of the primary unit. Or, the researcher could have applied both stratification criteria -one on the hills and another on the panicles-and the resulting sampling design would have been a two-stage sampling with stratification of both the primary and secondary units.

Sub-sampling with an auxiliary variable. The main features of a design for subsampling with an auxiliary variable are:

- In addition to the character of interest, say X, another character, say Z, which is closely associated with and is easier to measure than X, is chosen.
- Character Z is measured both on the main sampling unit and on the subunit, whereas variable X is measured only on the subunit. The subunit is smaller than the main sampling unit and is embedded in the main sampling unit.

This design is usually used when the character of interest, say X, is so variable that the large size of sampling unit or the large sample size required to achieve a reasonable degree of precision or both, would be impractical. To improve the precision in the measurement of X , without unduly increasing either the sample size or the size of sampling unit, the subsampling with an auxiliary variable design can be used.

Supplementary Techniques

So, far, we have discussed sampling techniques for individual plots, each of which is treated independently and without reference to other plots in the same experiment. However, in a replicated field trial where the sampling technique is to be applied to each and all plots in the trial, a question usually raised is whether the same set of random sample can be repeated in all plots or whether different random processes are needed for different plots. And, when data of a plant character are measured more than once over time, the question is whether the measurements should be made on the same samples at all stages of observation or should randomization be applied.

The two techniques aimed at answering these questions are **block sampling** and **sampling for repeated measurements.**

Block Sampling is a technique in which all plots of the same block (i.e. replication) are subjected to the same randomization scheme (i.e. using the same sample location in the plot) and different sampling schemes are applied separately and independently for different blocks. The block sampling technique has the following desirable features:

- Randomization is minimized. With block sampling randomization is done only r times instead of rt times as it is when randomization is done separately for each and all plots.
- Data collection is facilitated. With block sampling, all plots in the same block have the pattern of sample locations so that an observer (data collector) can easily move from plot to plot within a block without the need to reorient himself to a new pattern of sample location.
- Uniformity between plots of the same block is enhanced because there is no added variation due to changes in sample location from plot to plot.

Data collection by block is encouraged. For example, if data collection is to be done by several persons, each can be conveniently assigned to a particular block which facilitates the speed and uniformity of data collection. Even if there is only one observer for the whole experiment, he can complete the task one block at a time, taking advantage of the similar sample locations of plots in the same block and minimizing one source of variation among plots, namely, the time span in data collection.

Sampling for Repeated Measurements: Plant characters are commonly measured at different growth stages of the crop. For example, tiller number in rice may be measured at 30, 60, 90 and 120 days after transplanting or at the tillering, flowering, and harvesting stages. If such measurements are made on the same plants at all stages of observation, the resulting data may be biased because plants that are subjected to frequent handling may behave differently from others. In irrigated wetland rice, for example, frequent trampling around plants, or frequent handling of plants not only affect the plant characters being measured but also affect the plants' final yields. On the other hand, the use of an entirely different set of sample plants at different growth stages could introduce variation due to differences between sample plants. The partial replacement procedure provides for a satisfactory compromise between the two conflicting situations. With partial replacement, only a portion p of the sample plants used in one growth stage is retained for measurement in the succeeding stage. The other portion of (1-p) sample plants is randomly obtained from the remaining plants in the plot. The size of p depends on the size of the estimated undesirable effect of repeated measurements of the sample plants in a particular experiment. The smaller this effect, the larger p should be. For example, in the measurement of plant height and tiller number in transplanted rice, p is usually about 0.75. That is, about 75% of the sample plants measured at given growth stage is retained for measurement in the succeeding stage and the remaining 25% is obtained at random from the other plants in the plot.

Analysis

The various steps involved in the analysis of sampled data is described here considering a block design setting. Suppose an experiment is conducted with 't' treatments replicated 'r' times and let there be 'n' observations made in each plot. We assume the following linear additive model for the block design.

$$Y_{ijk} = \mu + \tau_i + \beta_j + e_{ij} + \eta_{ijk}$$

where Y_{ijk} is the observation on the kth sample for the ith treatment in the jth replicate (i = 1,2,...,t ; j = 1,2,...,r; k = 1,2,...,n), μ is the general mean effect, τ_i is the effect of ith treatment, β_j is the effect of jth replication, e_{ij} is the plot error distributed as $N(0 , \sigma_e^2)$, η_{ijk} is the sampling error distributed as $N(0 , \sigma_s^2)$.

The analysis of variance will be of the form given below:

ANOVA

Source	D.F.	S.S	M.S.	E(M.S.)
Replication	(r-1)	SST		
Treatments	(t-1)	SSR		$\sigma_s^2 + n\sigma_e^2 + \frac{rn}{t-1}\sum_j(\tau_i - \tau.)^2$
Treatment x Replication (Plot error)	(t-1) (r-1)	SSRT	s_1^2	$\sigma_s^2 + n\sigma_e^2$
Sampling Error (Samples within plots)	rt(n-1)	SSE	s_2^2	σ_s^2
Total	rtn-1			

The sampling error is estimated as $\hat{\sigma}_s^2 = s_2^2$.

The plot error is estimated as $\hat{\sigma}_e^2 = \frac{s_1^2 - s_2^2}{n}$.

When $\hat{\sigma}_e^2$ is negative, it is taken as zero.

The variance of the ith treatment mean ($\overline{Y}_{i..}$) based on r-replications and s-samples per plot = $\frac{\sigma_s^2 + n\sigma_e^2}{rn}$

The estimated variance of ($\overline{Y}_{i..}$) = $\frac{(\hat{\sigma}_s^2 + n\hat{\sigma}_e^2)}{rn}$

Taking the number of sampling units in a plot to be large (infinite), the estimated variance of a treatment mean when there is complete recording (i.e. the entire plot is harvested) = $\frac{\hat{\sigma}_e^2}{r}$

The efficiency of sampling as compared to complete recording

$$\frac{\hat{\sigma}_e^2 / r}{(\hat{\sigma}_s^2 + n\hat{\sigma}_e^2) / rn}$$

The standard error of a treatment mean ($\overline{Y}_{i..}$) with 'n' samples per plot and with 'r' replications is

$$\left[\frac{\hat{\sigma}_s^2}{rn}+\frac{\hat{\sigma}_e^2}{r}\right]^{1/2}$$

The percentage standard error or coefficient of variation is

$$p = \left[\left(\frac{\hat{\sigma}_s^2}{rn}+\frac{\hat{\sigma}_e^2}{r}\right)^{1/2}\Big/(\bar{Y}_{i..})\right]x\,100$$

Thus

$$n = \frac{\hat{\sigma}_s^2}{r}\left[\frac{1}{\frac{p^2(\bar{Y}_{i..})^2}{(100)^2}-\frac{\hat{\sigma}_e^2}{r}}\right]$$

For any given r and p, there will be t values for s corresponding to the t treatment means. The maximum s will ensure the estimation of any treatment mean with a standard error not exceeding p percent.

The sum of squares due to different components of ANOVA can be obtained as follows:

Form a two way table between replications and treatments, each cell figure being the total over all samples from a plot.

Grand Total (G.T.) = $\sum_i\sum_j\sum_k y_{ijk}$, Correction factor (C.F.)= $\frac{(G.T.)^2}{rtn}$

Total S.S. = $\sum_i\sum_j\left(\sum_k y_{ijk}\right)^2\Big/n-C.F.$

T_i = ith treatment total = $\sum_j\sum_k y_{ijk}$

R_j = jth replication total = $\sum_i\sum_k y_{ijk}$

Treatment S.S. = $\sum_i\frac{T_i^2}{rn}-C.F.$, Replication S.S. = $\sum_j\frac{R_j^2}{tn}-C.F$

Replication x Treatment S.S. = Total S.S. - Replication S.S -Treatment S.S.

Total S.S. of the entire data = $\sum_i\sum_j\sum_k y_{ijk}^2-C.F.$

S.S. due to sampling error = Total S.S. of the entire data - Replication S.S. -
Treatment S.S. - Replication x Treatment S.S.

Exercise: To study the effect of differences in the number of plants per hill on the growth of Maize crop, a randomized block design was laid out at the Agricultural College Farm, Poona. The treatments tried were A - one plant per hill, B - two plants per hill, C - three plants per hill, D - four plants per hill.

The net plot size used in the layout was 26’ x 20’ and the spacing between hills was 2’ x 2’. The table below gives the data on the length (in inches) of 5 cobs randomly selected from each plot:

Length of cobs (in inches)

Replication	Cob number	Treatments			
		A	B	C	D
I	1	9.3	9.0	8.6	6.4
	2	8.8	9.0	7.0	7.2
	3	9.0	10.5	8.4	6.8
	4	8.8	8.9	9.1	7.7
	5	8.6	9.2	8.2	6.0
II	1	10.2	9.7	9.0	6.4
	2	9.0	10.0	8.0	7.4
	3	9.4	9.2	8.1	6.8
	4	9.6	10.5	8.2	6.8
	5	9.8	10.3	7.0	6.6
III	1	9.9	8.4	7.5	6.3
	2	10.4	9.4	7.5	6.7
	3	11.0	8.2	8.5	6.0
	4	10.8	9.1	8.0	7.0
	5	10.0	9.8	8.6	7.3
IV	1	10.6	8.8	7.0	8.4
	2	9.2	9.3	7.3	7.8
	3	9.9	9.9	7.6	8.0
	4	10.4	9.0	6.7	8.4
	5	9.9	8.0	6.5	7.5
V	1	10.4	11.0	9.9	7.7
	2	9.0	10.4	9.0	7.0
	3	9.7	9.0	8.9	7.0
	4	9.3	10.2	8.9	6.7
	5	9.6	9.6	9.4	7.2

(a) Analyze the data and find the standard error of treatment means.

(b) Estimate the plot and sampling components of error variance and use these estimates to find out the relative efficiency of sampling.

(c) Prepare a table giving the minimum number of sampling units per plot necessary to estimate the treatment means with 4 and 5 percent standard error when the number of replications are 5 and 6.

Calculations

Step 1: Form the following two way table between replications and treatments, each cell figure being the total of cob lengths in five samples from a plot.

Replication	Treatments				Total
	A	B	C	D	
I	44.5	46.6	41.3	343.1	166.5
II	48.0	49.7	40.3	34.0	172.0
III	52.1	44.9	40.1	33.3	170.4
IV	50.0	45.0	35.1	40.1	170.2
V	48.0	50.2	46.1	35.6	179.9
Total	242.6	236.4	202.9	177.1	859.0

Step 2: Calculation of sum of squares and Analysis of variance.

The various sum of squares can be obtained using the formulae given above and the Analysis of Variance table can be obtained.

ANOVA

Source	**D.F.**	**S.S.**	**M.S.**	**F**
Replication	4	4.91	1.23	0.59
Treatment	3	112.09	37.36	18.05**
Replication x Treatment (plot error)	12	24.88	2.07	6.68*
Samples within plots (Sampling error)	80	24.91	0.31	
Total	99	166.79		

** denotes significant at 1 percent level and * significant at 5 percent level.

The mean square (s_1^2) is first tested against s_2^2 if - (i) s_1^2 is significant, then treatments are tested against s_1^2 and if -(ii) s_1^2 is not significant , the treatments are tested against the pooled mean square of s_1^2 and s_2^2.

In the present case s_1^2 is significant, so we test the treatments against s_1^2.

Step 3: Standard Error

Standard Error of the difference between two treatment means

$$\text{S.E}_d = \sqrt{\frac{s_1^2}{rn}} = \sqrt{\frac{2 \text{ x } 2.07}{5 \text{ x } 5}} = 0.4069 \text{ inches.}$$

Step 4: Efficiency

$$\hat{\sigma}_e^2 = \frac{s_1^2 - s_2^2}{n} = \frac{2.07 - 0.31}{5} = 0.3520$$

$$\hat{\sigma}_s^2 = s_2^2 = 0.31$$

The estimated variance of

$$\overline{Y}_{i..} = \frac{\hat{\sigma}_s^2}{rn} + \frac{\hat{\sigma}_e^2}{r} = \frac{2.070}{25} = 0.0828$$

Estimated variance in case of complete recording $= \frac{\hat{\sigma}_e^2}{r} = \frac{0.352}{5} = 0.0704.$

Efficiency of sampling as compared to complete recording

$$\frac{\hat{\sigma}_e^2 / r}{(\hat{\sigma}_s^2 + n\hat{\sigma}_e^2)/rn} = 0\ .85$$

Step 5: Estimation of sampling units per plot

$$s = \frac{\hat{\sigma}_s^2}{r}\left\{\frac{1}{\frac{p^2(\overline{Y}_{i..})^2}{(100)^2} - \frac{\hat{\sigma}_e^2}{r}}\right\}$$

Thus the number of sampling units required to measure the treatment means with 4 and 5 per cent standard error when the number of replication are 5 and 6 is worked out and is presented below.

Sampling units per plot (s)

Treatments	Treatment means	p=4		p=5	
		r = 5	r = 6	r = 5	r = 6
1	9.704	1	1	1	1
2	9.456	1	1	1	1
3	8.116	2	2	1	1
4	7.084	5	3	2	1

Step 6: Conclusion

(a) The treatments are found to be highly significant.

(b) Efficiency of sampling as compared to complete recording is 85 per cent.

(c) The number of sampling units necessary to estimate treatment means with

(i) 4 per cent standard error
when number of replications is 5 is 5,
when number of replications is 6 is 3.

(ii) 5 per cent standard error
when number of replications is 5 is 2,
when number of replications is 6 is 1.

PROBIT ANALYSIS OF DOSE-RESPONSE DATA

1. Introduction

Probit analysis is widely used in various fields where the response variable is qualitative. The main application of probit analysis is observed in the field of toxicological studies, where it transforms the sigmoid dose-response curve to a straight line that can then be easily analyzed by regression either through least squares or maximum likelihood. In other words, probit analysis is a methodology which transforms the complex percentage affected vs. dose response into a linear relation of probit vs. dose response. The probit can then be translated into percentages. The method is appropriate because of the typical shape found in the dose response curve. The method is approximate but allows quantification of consequence due to exposure.

"Probit" is an abbreviation of the term "probability unit" (the term is attributed to Bliss) and was the first such model developed and studied to treat data such as the percentage of pest killed by a pesticide. Bliss(1934) proposed transforming the percentage killed into "probit" (he defined it arbitrarily as equal to 0 for 0.0001 and 10 for 0.9999) and included a table to aid other researchers to convert kill-percentage to probit, which then could be plotted against the logarithm of the dose i.e. dosage. The table introduced by Bliss was carried forward in an important text on toxicological application by Finney (1952). Values tabulated by Bliss can be derived from probit as defined here by adding a value of 5. Using Bliss's idea, Leslie *et al*. (1945) were able to discuss the distribution of body–weight at which female rats in the wild reach maturity through probit analysis.

Mainly Probit analysis is used to analyze data from bioassays [most commonly refers to assessment of vitamins, hormones, toxicants and drugs of all kinds by means of response produced when doses are given to experimental animals (Finney 1952)] experiments, such as proportion of insect killed by several concentrations of an insecticide or at several time intervals at one or more concentration of an insecticide (Throne *et al*., 1995). One type of assay which has been found valuable in many different fields, but especially in toxicological studies, is that dependent upon quantal or all-or-nothing response. Though quantitative measurement of a response is almost always to be preferred when practicable, there are certain responses which permit no graduation and which can only be expressed as 'occurring' or 'not-occurring'. The most common example is mortality such as in many insecticidal studies the interest lies in whether or not a test insect is dead, or whether the insect has reached a certain degree of inactivation. In fungicidal investigations, failure of a spore to germinate is a quantal response of similar importance.

2. Probit Model

In probability theory and statistics, the probit function is the inverse cumulative distribution function (CDF), associated with the standard normal distribution. An alternative distribution could be the logistic distribution, which leads to the logit or logistic model. Both logistic and probit curves are so similar that they yield almost identical results. In practice they give estimated probabilities that differs very little (Aldrich and Nelson, 1984). The choice between logistic and probit is a matter of practical preference and experience.

For the standard normal distribution N (0, 1), the CDF is commonly denoted by Φ (z) (continuous, monotone increasing sigmoid function) given by,

$$\Phi(z) = \mathrm{P}(Z \le z) = \int_{-\infty}^{z} \varphi(u)\mathrm{du} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} \mathrm{du} \qquad \text{… (2.1)}$$

As an example, considering the familiar fact that the N (0, 1) distribution places 95% of probability between -1.96 and 1.96, and is symmetric about zero, it follows that

$$\Phi(-1.96) = 0.025 = 1 - \Phi(1.96) \qquad \text{… (2.2)}$$

The probit function gives the 'inverse' computation, generating a value of an N (0, 1) random variable, associated with specified cumulative probability. Formally, the probit function is the inverse of Φ (z), denoted by Φ^{-1}(p). Continuing the example,

$$\Phi^{-1}(0.025) = -1.96 = -\Phi^{-1}(0.975) \qquad \text{… (2.3)}$$

In general,

$$\Phi(\mathrm{probit(p)}) = \mathrm{p} \ \text{ and } \ \mathrm{probit}(\Phi(\mathrm{z})) = \mathrm{z} \qquad \text{… (2.4)}$$

In statistics, a probit model is a popular specification of a generalized linear model. If Y be a binary response variable, and let X be the single predictor variable, then the probit model assumes that,

$$\begin{aligned} \mathrm{P}(\mathrm{Y_i} = 1 | \mathrm{X_i} = \mathrm{x}) &= \Phi(\alpha + \beta \mathrm{x_i}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta \mathrm{x_i}} e^{-\frac{1}{2}z^2} \mathrm{dz} \end{aligned} \qquad \text{… (2.5)}$$

where Φ is the CDF of the standard normal distribution. The parameters β are estimated by maximum likelihood.

3. Quantal Response

3.1 Frequency Distribution of Tolerance

Two major components in any dose-response situation are the *stimulus* (e.g. a vitamin, a drug, a mental test or a physical force) and the *subject* (e.g. an animal, a plant, a human volunteer etc.). A stimulus is applied to the subject at a specified dose, intensity specified in units of concentration, weight, time or other appropriate measure, under controlled environmental condition. As a result subject manifests a response.

The response is quantal, occurrence or non-occurrence will depend upon the intensity of the stimulus. For any subject under controlled conditions, response occurs above a certain level of intensity, such a value is generally known as *threshold* or *limen*, but *tolerance* is now widely accepted. The tolerance value will vary from one subject to another in the population used. For quantal response data it is therefore necessary to consider distribution of tolerance over the population studied. If the dose or intensity of stimulus is measured by z, the distribution of tolerance may be expressed by:

$$dP = f(z)dz \qquad \dots (3.1)$$

This equation states the proportion, dP, of the whole population of subject whose tolerance lie between z and $z+dz$ at the time of testing, where dz represents a small interval on dose scale; the factor relating dP to the length of this interval is the frequency function, $f(z)$, uniquely determined for each possible value of z.

If a dose z_0 were given to the whole population, every individual whose tolerance was less than z_0 would respond. The proportion of these is P,
where

$$P=\int_0^{z_0} f(z)dz \qquad ... (3.2)$$

The measure of dose is here assumed to be a quantity that can conceivably range from zero to $+\infty$, response being certain for very high doses so that

$$\int_0^{\infty} f(z)dz = 1 \qquad \dots (3.3)$$

3.2 The Dose Metameter

The frequency distribution of tolerances, as measured on the natural scale, is usually markedly skewed, but often a simple transformation of the scale of measurement will convert it to a distribution approximately of normal form. The transformed scale of dose on which tolerances are normally distributed is known as *metametric* scale, and the measure of dose is the *dose* metameter.

The transformation

$$x = \log_{10} z \qquad \dots (3.4)$$

generally brings normality in the response variable, however for some fungicide a better transformation may be

$x = z^i$, where usually $i \leq 1$.

3.3 The Median Effective Dose

The effectiveness of a stimulus in relation to a quantal response is referred to as the *minimal effective dose,* or, for a more restricted class of stimuli as the *minimal lethal dose.* However it does not take into account the variation in tolerance within a population. The logical weakness of such concepts is the assumption that there is a dose for any given chemical, which is only just sufficient to kill all or most of the animals of a given species, and that doses a bit lesser would not kill any animal of that species. However, in toxicological studies such assumptions do not always hold good.

It might be thought that the minimal lethal dose of a poison could instead be defined as the dose just sufficient to kill a member of the species with the least possible tolerance, and also a *maximal non-lethal* dose as the dose, which will just fail to kill the most resistant member. Some doses will be so low that no test subject will succumb to them and others so high as to prove fatal at all and difficulties arise in determination of the end-points of these ranges. The problem is that of determining the dose at which the dose response curve for the whole population needs the 0% or 100% levels of kill and even a very large experiment could scarcely estimate these points with any accuracy.

Alternatively, a *median lethal dose*, or, as a more general term to include response other than mortality, a *median effective dose* is preferred. This is the dose that will produce a response in half the population. The median effective dose is commonly referred to as the ED_{50}, the more restricted concept of median lethal dose as the LD_{50}. With a fixed total number of subjects, effective doses in the neighborhood of ED_{50} can usually be estimated more precisely than those for more extreme percentage levels and this is, therefore, particularly favoured in expressing the effectiveness of the stimulus. The ED_{50} can be regarded as the median of the tolerance distribution and thus it is the level of tolerance such that exactly half the subject lies on either side of it.

For any distribution of tolerance, the ED_{50} is the value of z_0, such that

$$\int_0^{z_0} f(z)dz = 0.5 \qquad \ldots (3.5)$$

When a simple normalizing transformation for the doses is available, so that x, the normalizing measure of dose (commonly known as dosage), has a normally distributed tolerance, equation (3.1) is transformable to

$$dP = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx\,. \qquad \ldots (3.6)$$

where μ is the center of the distribution and σ^2, its variance. The μ is the population value of the mean dosage tolerance, or median effective dosage, and efforts must be directed at estimating it from the observational data. The $\log_{10}ED_{50}$ is the value of x_0 for which

$$\int_{-\infty}^{x_0} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 0.5 \qquad \ldots (3.7)$$

The solution of equation (3.7) is μ, so that the ED_{50} is 10^{μ}.

Any two insecticides may require the same rate of application in order to be effective to half the population, but, if the distribution of tolerances has a lesser 'spread' for one than for the other, any increase or decrease from this rate will produce a greater change in mortality for the first than for the second. This spread is measured by the variance σ^2. This measure along with the ED_{50} fully describes the effectiveness of the stimulus. The smaller the value of σ^2, the greater is the effect on mortality of any change in dose.

4. Estimation of the Median Effective Dose

4.1 The N.E.D. and Probit Transformation

Initially the measure of the probability of response was proposed on a transformed scale i.e. the normal equivalent deviate (or N.E.D.). This response metameter is Y, defined by:

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y} e^{-\frac{1}{2}u^2} du \qquad \text{... (4.1)}$$

Thus the N.E.D. of any value of P between 0 and 1 is defined as the abscissa corresponding to a probability P in a normal distribution with mean 0 and variance 1.

Equation (4.1) determines either of P and Y uniquely from the other. From integration of equation (3.6), if P is the probability of response at a dosage whose metameter is a particular value X, then

$$P = \int_{-\infty}^{X} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \qquad \text{… (4.2)}$$

which by writing x = μ+σu

$$\text{becomes} \quad P = \int_{-\infty}^{\frac{(X-\mu)}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \qquad \text{… (4.3)}$$

Comparison of equation (4.3) with equation (4.1) shows that

$$Y = \frac{(X - \mu)}{\sigma} \qquad \text{… (4.4)}$$

Thus, the relation between the dose metameter (x) and the N.E.D. of the probability of response at that dosage is a straight line.

Bliss (1934) suggested a slightly different response metameter. Bliss defined the probit of the proportion P as Y, where

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Y-5} e^{-\frac{1}{2}u^2} du \qquad \text{… (4.5)}$$

For any P, the probit is simply the N.E.D. increased by 5. All subsequent theory is essentially same for the two metameters. The N.E.D, however, is negative if P is less than 50%, whereas the probit is generally positive unless P is exceedingly small.

Comparison with equation (4.1) shows that the probit of the expected proportion responded is related by the linear equation

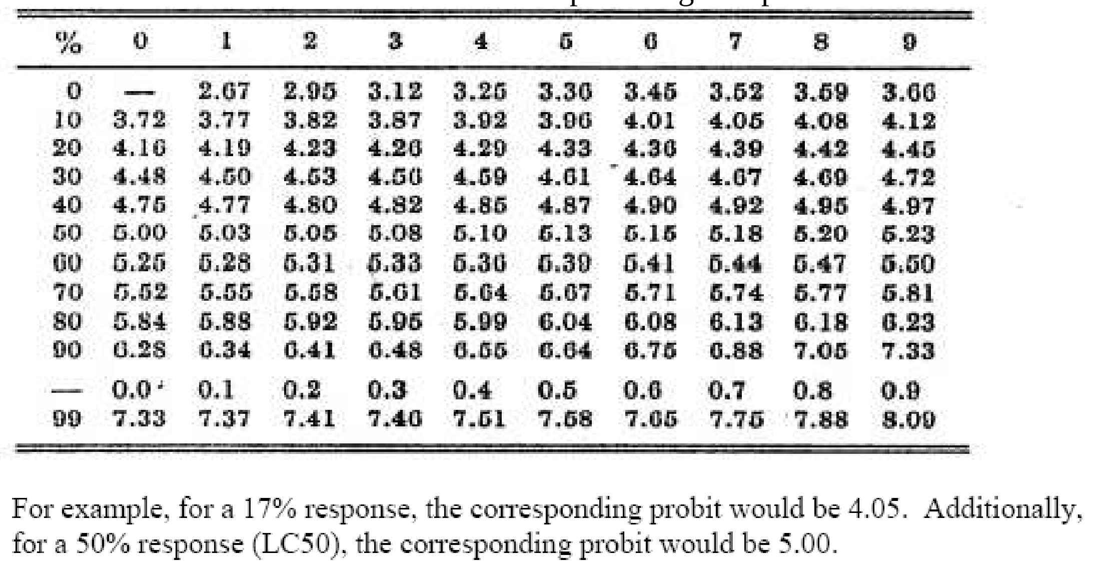
$$Y = 5 + \frac{1}{\sigma}(x-\mu) \qquad \text{… (4.6)}$$

In particular, the median effective dosage is estimated as that value of *x* which gives Y = 5.

4.2 The Probit Regression Line When experimental data on the relationship between dosage and response have been obtained, either a graphical or an arithmetical approach can be used to estimate the parameters. Both approaches depend on the probit transformation. The graphical approach is much more rapid and is sufficiently good for many purposes, but for some, more complex problems, or when an accurate assessment of the precision of estimates is wanted, the more detailed arithmetical analysis is necessary. Here graphical approach is discussed.

To start with, the percentage response observed for each dose are calculated and converted to probits by means of the following table (Finney, 1971):

Table 4.1: Transformation of percentages to probits



%	0	1	2	3	4	5	6	7	8	9
0	—	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
—	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

For example, for a 17% response, the corresponding probit would be 4.05. Additionally, for a 50% response (LC50), the corresponding probit would be 5.00.

The probits are then plotted against the dose metameter i.e. the logarithm (base 10) of the dose. Very extreme probits, say outside the range 2.5-7.5, carry little weight and should be disregarded. A straight line is drawn to fit the points as satisfactorily as possible. The line is nothing but the weighted regression line of the mortality probit on x. By visual inspection, the $\log_{10}ED_{50}$ is estimated from the line as m, the dosage at which Y = 5. This can be taken as estimate of μ.The estimated slope of the line (b) is an estimate of $1/\sigma$, can be obtained as the increase in Y for a unit increase in x. These two estimates are then substituted for the parameters in equation (4.6) to give the estimated relation between dosage and response. To test the hypothesis that the line is an adequate representation of the data, a χ^2 test of the form

$$\chi^2 = \sum \frac{(r-np)^2}{np(1-p)} \sim \chi^2_{k-2} \qquad \text{… (4.7)}$$

may be used. Here n is the number of subjects exposed to a specific concentration, r is the observed number of units respond out of n number of unit, $p = \frac{r}{n}$ is the estimated proportion of response for that particular concentration. Here k level of concentration is applied over the test subject and summation is taken over all the level of concentration tested. A value of χ^2 within the limits of random variation indicates satisfactory agreement theory (the line) and observation (the data).

Example 4.1: Table 4.2 contains the data on effect of a series of concentrations of the pesticide Rotenone when spraying on *Macrosiphoniella sanborni*, the chrysanthemum aphis, in batches of about fifty (Finney, 1971).

Table 4.2: Toxicity of Rotenone to *Macrosiphoniella sanborni*

Concentration (mg. /1.)	**No. of insects (n)**	**No. of affected (r)**	**% kill (p)**	**Log concentration (x)**	**Empirical probit**
10.2	50	44	88	1.01	6.18
7.7	49	42	86	0.89	6.08
5.1	46	24	52	0.71	5.05
3.8	48	16	33	0.58	4.56
2.6	50	6	12	0.41	3.82
0	49	0	0	-	-

Table 4.2 summarizes the dose metameter, percentage kill, and empirical probit values for the experiment. Over the range of concentrations tested, the relation between percentage kills and log concentration is apparently sigmoid. The percentages are plotted against the logarithm of doses and fitted with the normal sigmoid curve in Fig. 4.1.

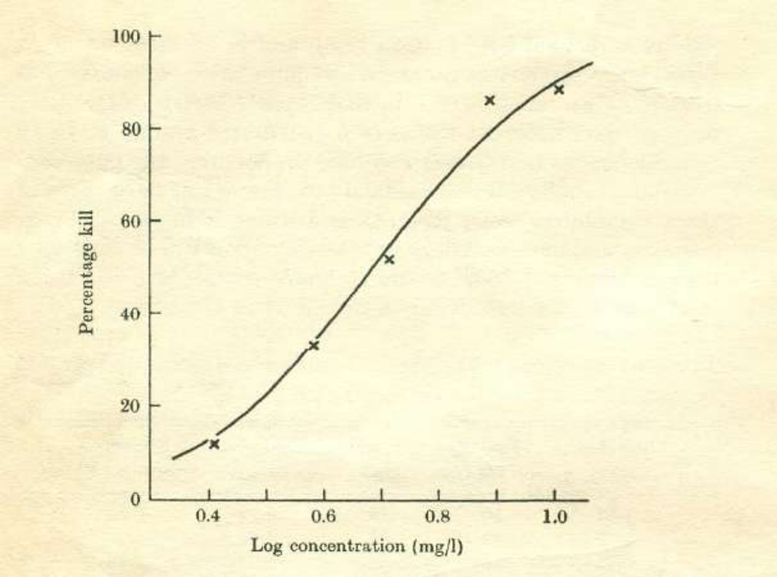


Fig.4.1: Relation between percentage kill of ***Macrosiphoniella sanborni*** and logarithm of dose of Rotenone.

In order to fit a straight line, percentages of kill have been converted into probits using Table 4.1 and are given in the last column of Table 4.3. When probits are plotted against dosages (logarithm to the base 10 of doses); they lie nearly on a straight line. Fig 4.2 gives the plot of probits vs. dosages. From this line, probits corresponding to many different values of x can be found out and converted back to percentages by using Table 4.1 inversely.

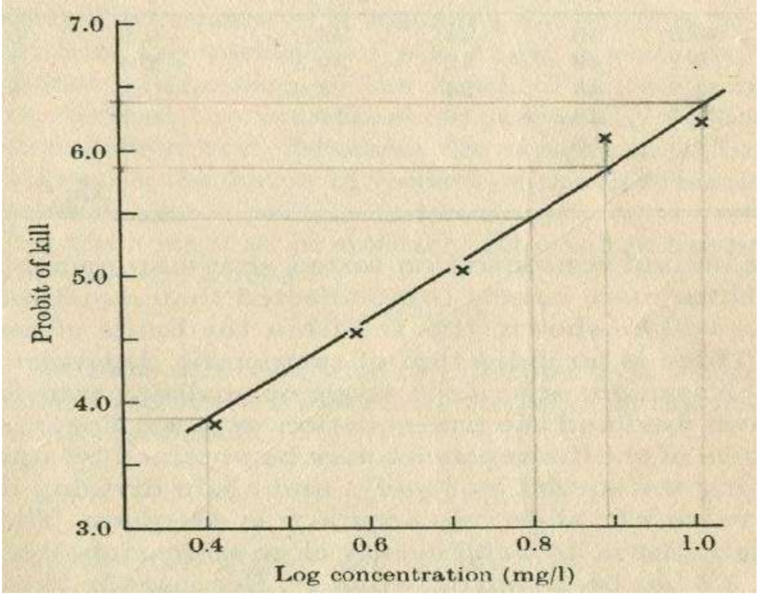


Fig. 4.2: Relation between probit of kill of ***Macrosiphoniella sanborni*** and logarithm of dose of Rotenone.

In Fig. 4.2 of the present example, a probit value of 5 is given by a dosage of m = 0.687; this therefore is the estimate of $\log_{10}ED_{50}$, and the ED_{50} is estimated as a concentration of 4.86mg/l. Similarly the $\log_{10}ED_{90}$ corresponds to a probit of 6.28 and is therefore 1.006; the ED_{90} is thus estimated as 10.14 mg/l.

Thus Fig. 5.2 can also be used to give the slope of the line: an increase of 0.319 in x corresponds with an increase of 1.28 in probit. Hence the estimated regression coefficient of probit on dosage, or the rate of increase of probit value per unit increase in x, is

$$b = 4.01 \qquad \ldots (4.8)$$

Thus equation (4.6) becomes

$$Y = 5 + 4.01\,(x - 0.687), \text{ or } Y = 2.25 + 4.01x \qquad \ldots (4.9)$$

Equation (4.9) may be used to calculate expected numbers of insects killed at each concentration. By substitution of the values of x used in the experiment, the equation gives the values of Y which are given in column 2 of Table 4.3 as expected probits. Thus a probit of 6.30 corresponds to a percentage of between 90 and 91, or, more accurately, 90 + 2/6%. If the expected proportion for any concentration is multiplied by n, the number of insects tested at that concentration, the result is the expected number of responded insects, or the average number which would be affected in a batch of size n if equation (4.9) represents the true relationship between dosage and response. These numbers, np, may then be compared with the actual numbers affected, r, in order to judge the adequacy of the equation.

Table 4.3: Comparison of Observed and Expected Mortality

Log concentration	**Expected probit**	**% kill**	**No. of insects**	**No. affected**		**Discrepancy**	$\frac{(r-np)^2}{np(1-p)}$
				Observed	Expected		
(x)	**(Y)**	**(p)**	**(n)**	(r)	(np)	**(r-np)**	
1.01	6.30	90.3	50	44	45.2	-1.2	0.33
0.89	5.83	79.7	40	42	39.1	2.9	1.06
0.71	5.10	54.0	46	24	24.8	-0.8	0.06
0.58	4.58	33.7	48	16	16.2	-0.2	0.00
0.41	3.90	13.6	50	6	6.8	-0.8	0.11
							$\chi^2_{[3]}$=1.56

Since proportion of response has been estimated from the data, the degree of freedom of χ^2(=3) is two less than the number of concentrations tested. From Fisher and Yates Table (1964, Table IV), the tabulated value of $\chi^2_{[3]}$ at 5% level of significance is 7.815. Thus the calculated value of $\chi^2_{[3]}$ (1.56) is much smaller than the tabulated value of $\chi^2_{[3]}$ at 5% level of significance. Hence, the probit regression line is very satisfactory representation of the results of the experiment.

5. Conclusions

Probit analysis has been widely used in diverse fields wherein the response variable is qualitative. Probit analysis for dose-response studies under regression framework is commonly done. In such studies, the estimation of the median effective dose (ED_{50}) i.e. the dose that will produce a response in half the population along with its variance can be chiefly done. This can be easily be achieved by using any standard statistical software.

LOGISTIC REGRESSION

1. Introduction

Regression analysis is a widely used method for obtaining a functional relationship between the response or dependent variable and one or more explanatory or predictor variables. In all the regression models, we implicitly assumed that the response variable is quantitative in nature whereas the explanatory variables are either quantitative, qualitative or a mixture thereof. In case of qualitative or non-metric response variable usual assumptions of regression models are violated, hence, it is better to look for alternative models. In practice, situations involving categorical outcomes are quite common. Suppose we want to study the labour force participation (LFP) decision of adult males. Since an adult is either in the labour force or not, LFP is a yes or no decision. Similarly, in the setting of evaluating an extension program, for example, predictions may be made for the dichotomous outcome of success/failure or improved/not-improved. An economist may be interested in determining the probability that an agro-based industry will fail given a number of financial ratios and the size of the firm (i.e. large or small) etc.

Usually discriminant analysis could be used for addressing each of the above problems. However, because the independent variables are mixture of categorical and continuous variables, the multivariate normality assumption may not hold. In these cases the most preferable technique is the logistic regression analysis as it does not make any assumptions about the distribution of the independent variables.

2. Violation of Assumptions of Linear Regression Model when Response is Qualitative

Linear regression is considered in order to explain the constraints in using such model when the response variable is qualitative. Consider the following simple linear regression model with single predictor variable and a binary response variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ , } i = 1, 2, \ldots, n$$

where the outcome Y_i is binary (taking values 0,1), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and are independent and n is the number of observations.

Let π_i denote the probability that Y_i =1 when X_i = x, i.e.

$$\pi_i = P(Y_i = 1 | X_i = x) = P(Y_i = 1)$$

thus $\quad P(Y_i = 0) = 1 - \pi_i$.

Under the assumption $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(Y_i) = 1.(\pi_i) + 0.(1 - \pi_i) = \pi_i$$

If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = 1 - \pi_i \qquad \text{when } Y_i = 1$$

$$\varepsilon_i = -\pi_i \qquad \text{when } Y_i = 0$$

Because the error is dichotomous (discrete), so normality assumption is violated. Moreover, the error variance is given by:

$$\begin{aligned} V(\varepsilon_i) &= \pi_i (1 - \pi_i)^2 + (1 - \pi_i)(-\pi_i)^2 \\ &= \pi_i (1 - \pi_i) \end{aligned}$$

It can be seen that variance is a function of π_i's and it is not constant. Therefore the assumption of homoscadasticity (equal variance) does not hold.

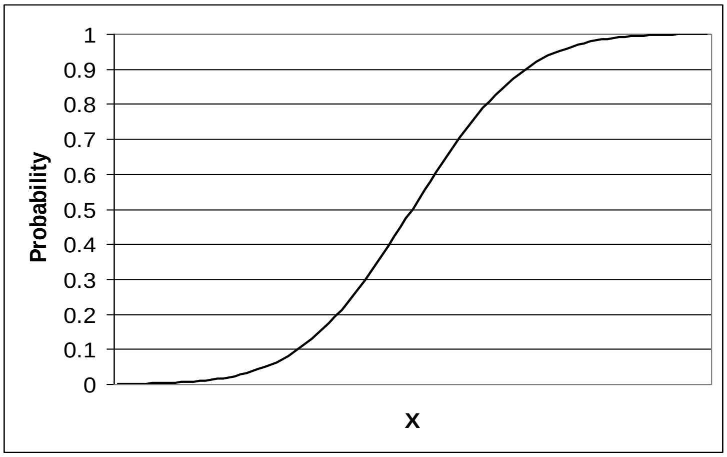
3. Logistic Regression

Logistic regression is normally recommended when the independent variables do not satisfy the multivariate normality assumption and at the same time the response variable is qualitative. Situations where the response variable is qualitative and independent variables are mixture of categorical and continuous variables, are quite common and occur extensively in statistical applications in agriculture, medical science etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model, developed primarily by a researcher named Cox during the late 1950s. Processes producing sigmoidal or elongated S-shaped curves are quite common in agricultural data. Logistic regression models are more appropriate when response variable is qualitative and a non-linear relationship can be established between the response variable and the qualitative and quantitative factors affecting it. It addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors. In logistic regression model, the predictors need not have to be normally distributed, the relationship between response and predictors need not be linear or the observations need not have equal variance in each group etc. A good account on logistic regression can be found in Fox (1984) and Kleinbaum (1994).

The problem of non-normality and heteroscadasticity leads to the non applicability of least square estimation for the linear probability model. Weighted least square estimation, when used as an alternative, can cause the fitted values not constrained to the interval (0, 1) and therefore cannot be interpreted as probabilities. Moreover, some of the error variance may come out to be negative. One solution to this problem is simply to constrain the probability of outcome to the unit interval while retaining the linear relation between probability of outcome and regressor within the interval. However, this constrained linear probability model has certain unattractive features such as abrupt changes in slope at the extremes 0 and 1 making it hard for fitting the same on data. A smoother relation between the probability of outcome and regressor is generally more sensible. To correct this problem, a positive monotone (i.e. non-decreasing) function is required to transform linear combination of regressor to unit interval. Any cumulative probability distribution function (CDF) meets this requirement. That is, re-specify the model as $\pi_i = P(\beta_0 + \beta_1 x_i)$. where, π_i is the probability of outcome and P is the cumulative distribution function. Moreover, it is advantageous if P is strictly increasing, for then, the transformation is one-to-one, so that model can be rewritten as $P^{-1}(\pi_i) = (\beta_0 + \beta_1 x_i)$, where P^{-1} is the inverse of the CDF P. Thus the non-linear model for itself will become both smooth and symmetric, approaching $\pi = 0$ and $\pi = 1$ as asymptotes. Thereafter maximum likelihood method of estimation can be employed for model fitting.

3.1 Properties of Logistic Regression Model

The Logistic response function resembles an S-shape curve, a sketch of which is given in the following figure. Here the probability π initially increases slowly with increase in X, and then the increase accelerates, finally stabilizes, but does not increase beyond 1.



The shape of the S-curve can be reproduced if the probabilities can be modeled with only one predictor variable as follows:

$\pi = P(Y=1|X=x) = 1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$\pi = P(Y=1|X_1 = x_1 ... X_r = x_r)$

$=1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$.

This equation is called the logistic regression equation. It is nonlinear in the parameters β_0, β_1… β_r. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

To explain the popularity of logistic regression, let us consider the mathematical form on which the logistic model is based. This function, called f (z), is given by

$f(z) = 1/(1+e^{-z})$, $-\infty < z < \infty$

Now when $z = -\infty$, $f(z) = 0$ and when $z = \infty$, $f(z) = 1$. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides

- Estimates that lie in the range between zero and one.
- An appealing S-shaped description of the combined effect of several explanatory variables on the probability of an event.

3.2. Maximum Likelihood Method of Estimation of Logistic Regression

Generally, the maximum likelihood method is used for estimating the parameters of the logistic regression model. The maximum likelihood estimates β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function. No closed-form solution exists for the values of β_0 and β_1 that maximize the log-likelihood function. Computer intensive numerical search procedures are therefore required to find the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Standard statistical software such as SPSS (Analyze- Regression-Binary

Logistic) provide maximum likelihood estimates for logistic regression. Once these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found, by substituting these values into the response function the fitted response function, say, $\hat{\pi}_i$, can be obtained. The fitted response function is as follows:

$$\hat{\pi}_i = \left(\frac{1}{1+e^{-\left(\hat{\beta}_0+\hat{\beta}_1 X_i\right)}} \right)$$

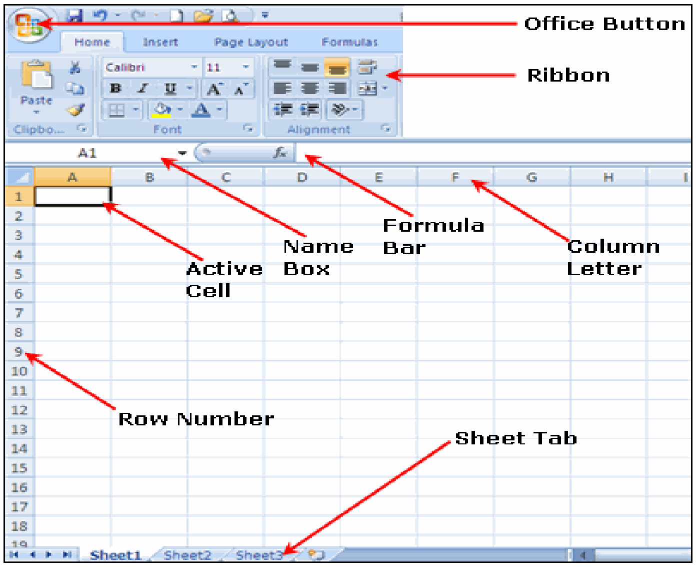
When log of the odds of occurrence of any event is considered using a logistic regression model, it becomes a case of logit analysis. Thus formed logit model will have its right hand side as a linear regression equation.

4. Practical Constraint

Sometimes quantitative information on adoption of a technology is not available but is available in qualitative form such as adopted / non-adopted, low / high adoption etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model. It can be used to describe the relationship of several independent variables to the binary (say, named 0 & 1) dependent variable. The logistic regression is used for obtaining probabilities of occurrence, say E, of the different categories when the model is of the form: $P(E = 1) = \frac{1}{1+\exp(-z)}$ where z is a function of associated variables, if $P(E = 1) \geq 0.5$ then there is more chance of occurrence of an event and if $P(E = 1) < 0.5$ then probability of occurrence of the event is minimum. If the experimenter wants to be more stringent, then the cutoff value of 0.5 could be increased to, say, 0.7.

MS-EXCEL: STATISTICAL PROCEDURES

Microsoft (MS) Excel () is a powerful spreadsheet that is easy to use and allows you to store, manipulate, analyze, and visualize data. It also supports databases, graphic and presentation features. It is a powerful research tool and needs a minimum of teaching. Spreadsheets offer the potential to bring the real numerical work alive and make statistics enjoyable. But the main disadvantage is that some advanced statistical functions are not available and it takes a longer computing time as compared to other specialized software.



Data Entry in Spreadsheets

- Data entry should be started soon after data collection in the field
- The raw data collected should be entered directly into computer. Calculations (e.g. % dry matter) or conversions (e.g. kg/ha to t/ha) by hand will very likely result in errors and therefore require more data checking once the data are in MS-Excel. Calculations can be written in MS-Excel using formulae (e.g. sum of wood biomass and leaf biomass to give total biomass).

Data Checking

One can use calculations and conversions for data checking. For example, if the collected data is grain yield per plot it may be difficult to see whether the values are reasonable. However, if these are converted to yield per hectare then one can compare the numbers with our scientific knowledge of grain yields. Simple formulae can be written to check for consistency in the data. For example, if tree height is measured 3 times in the year, a simple formula that subtracts 'tree height 1' from 'tree height 2'can be used to check the correctness of the data. The numbers in the resulting column should all be positive. We cannot have a shrinking tree! For new columns of calculated or converted data suitable header information (what the new column is, units and short name) at the top of the data should be included.

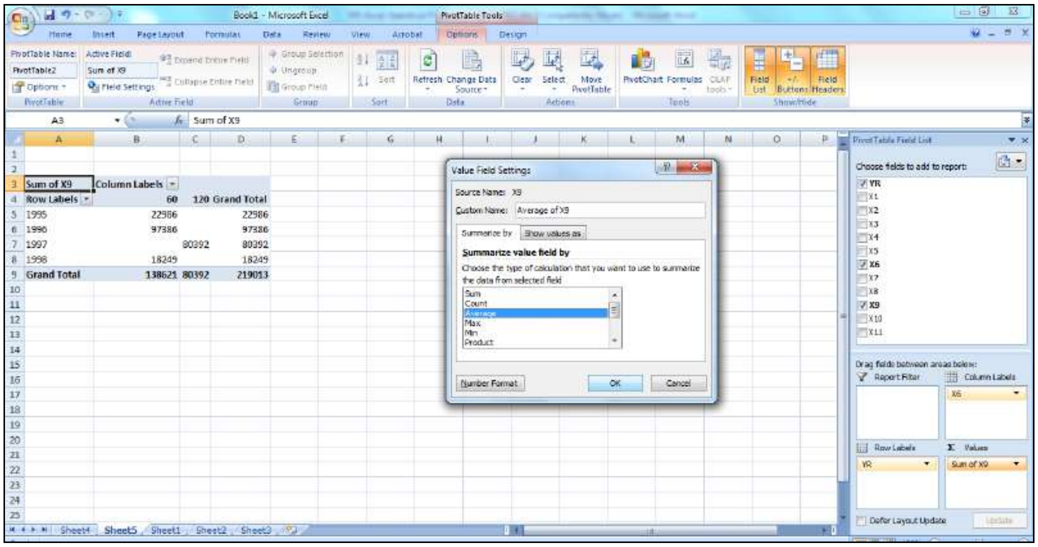
Missing Values

In MS-Excel the missing values are BLANK cells. It is useful to know this when calculating formulae and summaries of the data. For example, when calculating the average of a number of cells, if one cell is blank MS-Excel ignores this as an observation (i.e., the average is the sum/number of non-blank cells). But if the cell contains a '0' then this is included in the calculation (i.e., the average is the sum/no. of cells). In a column of 'number of fruit per plot', a missing value could signify zero (tree is there but no fruit), dead (tree was there but died so no fruit), lost (measurement was lost, illegible.) or not representative (tree had been browsed severely by goats). In this example, depending on the objectives of the trial, the scientist might choose to put a '0' in the cells of trees with no fruit and leave blank (but add comments) for the other 'missing values'.

Pivot Tables (to check consistency between replicates)

Variation between replicates is expected, but some level of consistency is also usual. We can use pivot tables to look at the data. A pivot table is an interactive worksheet table that quickly summarizes large amount of data using a format and calculation methods you choose. It is called pivot table because you can rotate its row and column heading around the core data area to give you different views of the source data. A pivot table provides an easy way for you to display and analyze summary information about data already created in MS-Excel or other application.

- Keep the cursor anywhere within the data range
- Choose “Insert” “Pivot Table” then “OK”
- From the “Pivot table Field List” drag and drop the respective fields under “Column Labels” , “Row Labels” and “Σ Values”
- Select “Value Field Settings” by clicking on the down arrow in “Σ Values” and choose the appropriate option and then click “OK”



Sum of X9	Column Labels		
Row Labels	60	120	Grand Total
1995	22986		22986
1996	97386		97386
1997		80392	80392
1998	18249		18249
Grand Total	138621	80392	219013

Value Field Settings

Source Name: X9

Custom Name: Average of X9

Summarize value field by: Average

Report Filter: (empty); Column Labels: X6; Row Labels: YR; Σ Values: Sum of X9

Scatter Plots (to check consistency between variates)

We can often expect two measured variables to have a fairly consistent relationship with each other. For example, 'number of fruits' with 'weight of fruits' or Stover yield plotted against grain yield. To look for odd values we could plot one against the other in a scatter plot. Scatter plots are useful tools for helping to spot outliers. This option is available under “Insert” menu.

Line Plots (to examine changes over time)

Where measurements on a 'unit' are taken on several occasions over a period of time it may be possible to check that the changes are realistic. A check back at the problematic data which is not in the usual trend can be made. . This option is available under “Insert” menu.

Double Data Entry

One effective, although not always practical, way of checking for errors caused by data entry mistakes is double entry. The data are entered by two individuals onto separate sheets that have the same design structure. The sheets are then compared and any inconsistencies are checked with the original data. It is assumed that the two data entry operators will not make the same errors. There is no 'built-in' system for double entry in MS-Excel. However, there are some functions that can be used to compare the two copies. An example is the DELTA function that compares two values and returns a 1 if they are the same and a 0 otherwise. To use this function we would set up a third worksheet and input a formula into each cell that compares the two identical cells in the other two worksheets. The 0's on the third worksheet will therefore identify the contradictions between the two sets of data. This method can also be used to check survey data but for the process to work the records must be entered in exactly the same order in both sheets. If a section at the bottom of the third worksheet contains mostly 0's, this could indicate that you have omitted a record in one of the other sheets.

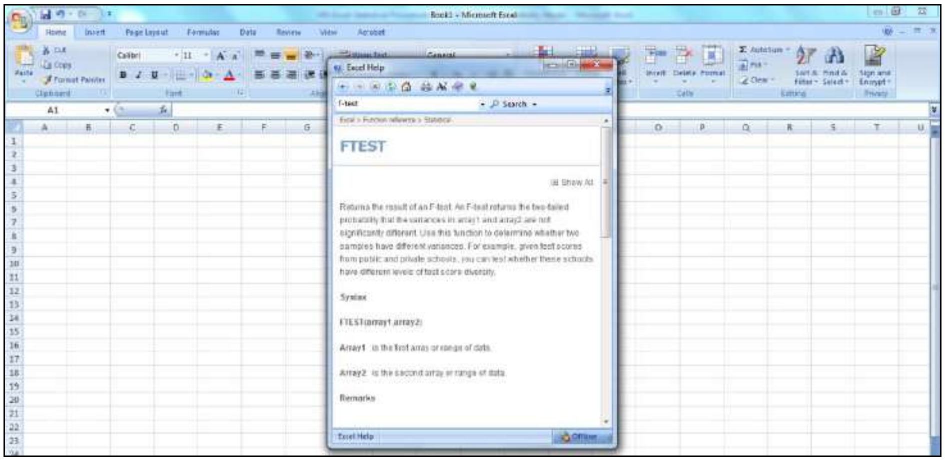
Preparing Data for Export to a Statistical Package

Statistical analysis of research data usually involves exporting the data into a statistical package such as GENSTAT, SAS or SPSS. These packages require you to give the MS-Excel cell range from which data are to be taken. In the latest editions of MS-Excel we can mark these ranges within MS-Excel and then transfer them directly into the statistical packages.

- Highlight the data you require including the column titles (the codes which have been used to label the factors and variables).
- Go to the Name Box, an empty white box at the top left of the spreadsheet. Click in this box and type a name for the highlighted range (e.g., Data). Press Enter.
- From now on, when you want to select your data to export go to the Name Box and select that name (e.g. Data). The relevant data will then be highlighted.

MS-Excel Help

If you get stuck on any aspect of MS-Excel then use the Help facility by clicking “F1” key. It contains extensive topics and by typing in a question you can extract the required information. See the snapshot below for an example:



FTEST

Returns the result of an F-test. An F-test returns the two-tailed probability that the variances in array1 and array2 are not significantly different. Use this function to determine whether two samples have different variances. For example, given test scores from public and private schools, you can test whether these schools have different levels of test score diversity.

Syntax

FTEST(array1,array2)

Array1 is the first array or range of data.

Array2 is the second array or range of data.

Remarks

Features of MS-Excel

Analytic Features

- The windows interface includes windows, pull down menus, dialog boxes and mouse support
- Repetitive tasks can be automated with MS-Excel. Easy to use macros and user defined functions
- Full featured graphing and charting facilities
- Supports on screen databases with querying, extracting and sorting functions
- Permits the user to add, edit, delete and find database records

Presentation Features

- Individual cells and chart text can be formatted to any font and font size
- Variations in font size, style and alignment control can be determined
- The user can add legends, text, pattern, scaling and symbols to charts.

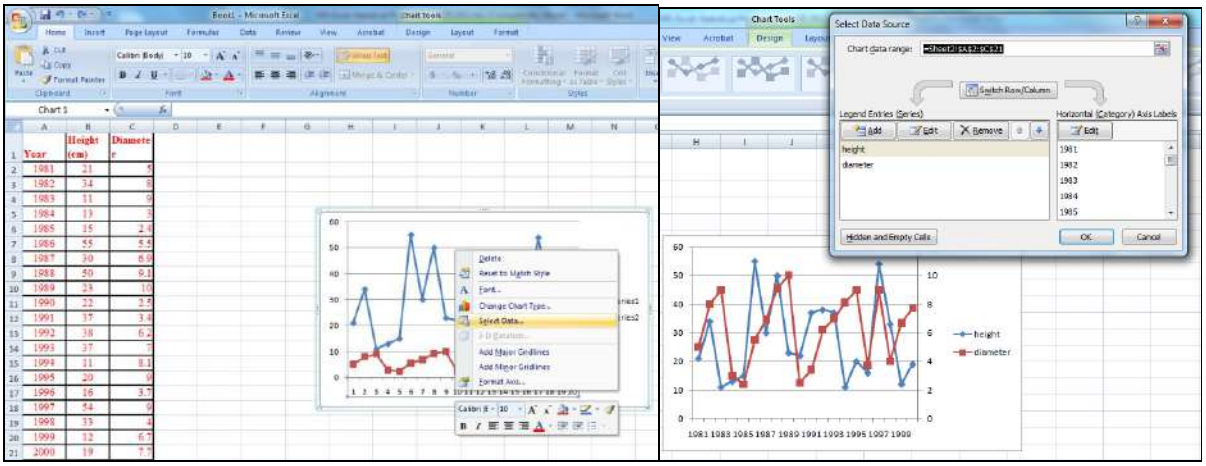
Charts and Graphs

A chart is a graphic representation of worksheet data. The dimension of a chart depends upon the range of the data selected. Charts are created on a worksheet or as a separate document that is saved with an extension xlsx. MS-Excel automatically scales the axes, creates columns categories and labels the columns. Values from worksheet cells or data points are displayed as bars, lines, columns, pie slices, or other shapes in the chart. Showing a data in a chart can make it clearer, interesting and easier to understand. Charts can also help the user to evaluate his/her data and make comparisons between different worksheet values.

Creating Line Chart

- Select relevant part of data
- Choose “Insert” “line”
- Select an appropriate option of line chart and click

Necessary changes in the chart can be done by clicking the right button of the mouse and choosing appropriate options.



Year	Height (cm)	Diameter
1981	21	5
1982	34	8
1983	11	9
1984	13	3
1985	15	2.4
1986	55	5.5
1987	30	6.9
1988	50	9.1
1989	23	10
1990	22	2.5
1991	37	3.4
1992	38	6.2
1993	37	7
1994	11	8.1
1995	20	9
1996	16	3.7
1997	54	9
1998	33	4
1999	12	6.7
2000	19	7.7

Chart data range: =Sheet2!$A$2:$C$21

Sorting and Filtering

MS-Excel makes it easy to organize, find and create report from data stored in a list.

Sort: To organize data in a list alphabetically, numerically or chronologically.

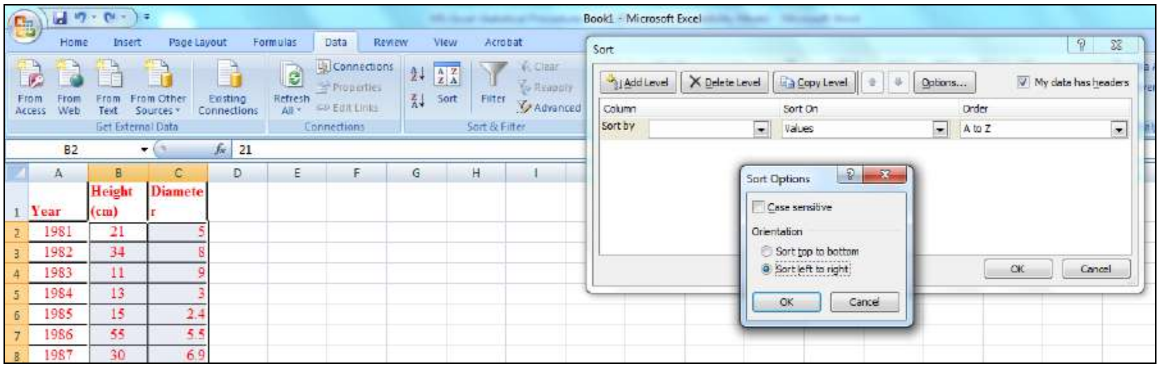
(i) To sort entire list

- Select a single cell in the list

- Choose “data” “sort”

(ii) Sorting column from left to right

- Choose the “option” button in the sort dialog box
- In the sort option dialog box, select “sort left to right”
- Choose “OK”



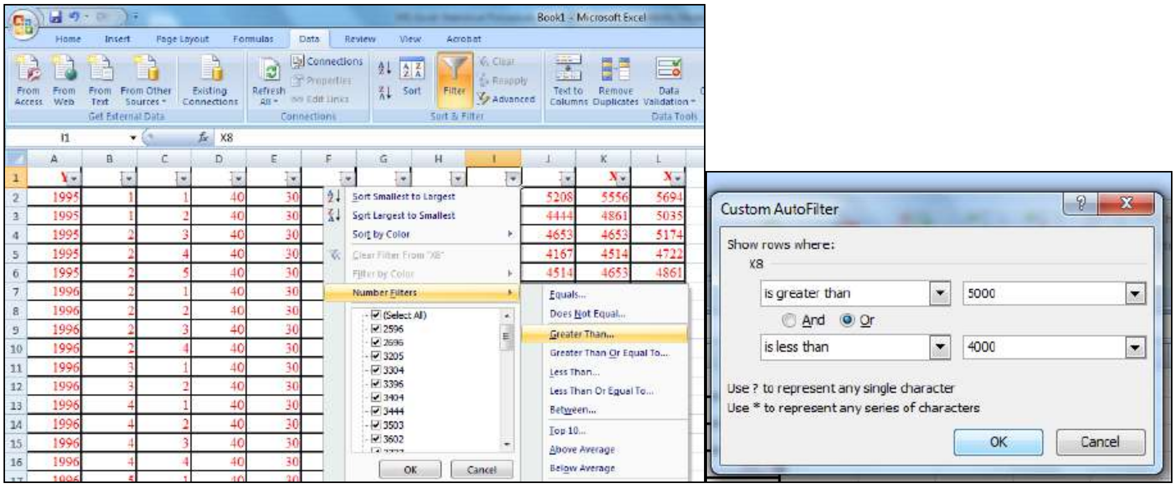
	A	B	C
1	Year	Height (cm)	Diameter
2	1981	21	5
3	1982	34	8
4	1983	11	9
5	1984	13	3
6	1985	15	2.4
7	1986	55	5.5
8	1987	30	6.9

Sort: ☑ My data has headers; Column: Sort by; Sort On: Values; Order: A to Z

Sort Options: □ Case sensitive; Orientation: ○ Sort top to bottom ● Sort left to right

Filter: To quickly find and work with a subset of your data without moving or sorting it.

- Choose “Data” and click on “Filter”
- MS-Excel place a drop down arrow directly on the column labels of the list
- Choose the column based on which the data has to be filtered. Clicking on the arrow displays a list of all the unique items in the column. Choose “Number Filter” option and define the required conditions.



Custom AutoFilter

Show rows where:
X8

is greater than | 5000

○ And ● Or

is less than | 4000

Use ? to represent any single character
Use * to represent any series of characters

Statistical Functions

Excel’s statistical functions are quite powerful. In general, statistical functions take lists as arguments rather than single numerical values or text. A list could be a group of numbers separated by commas, such as (3,5,1,12,15,16), or a specified range of cells, such as (A1:A6), which is the equivalent of typing out the list (A1,A2,A3,A4,A5,A6). The function COUNT(list) counts the number of values in a list, ignoring empty or nonnumeric cells, whereas COUNTA(list) counts the number of values in the list that have any entry at all. MIN(list) returns a list’s smallest value, whereas MAX(list) returns a list’s largest value. The functions AVERAGE(list), MEDIAN(list), MODE(list), STDEV(list) all carry out the statistical operations you would expect (STDEV stands for standard deviation), when you pass a list of values as an argument.

Create a Formula

Formulas are equations that perform calculations on values in your worksheet. A formula starts with an equal sign (=). For example, the following formula multiplies 2 by 3 and then adds 5 to the result: =5+2*3. The following formulas contain operators and constants:

Example formula	What it does
=128+345	Adds 128 and 345
=5^2	Squares 5

- Click the cell in which you want to enter the formula.
- Type **=** (an equal sign).
- Enter the formula.
- Press ENTER.

Create a Formula that Contains References or Names: A1+23

The following formulas contain relative references and names of other cells. The cell that contains the formula is known as a dependent cell when its value depends on the values in other cells. For example, cell B2 is a dependent cell if it contains the formula =C2.

Example formula	What it does
=C2	Uses the value in the cell C2
=Sheet2!B2	Uses the value in cell B2 on Sheet2
=Asset-Liability	Subtracts a cell named Liability from a cell named Asset

- Click the cell in which the formula enter has to be entered.
- In the formula bar, type **=** (equal sign).
- To create a reference, select a cell, a range of cells, a location in another worksheet, or a location in another workbook. One can drag the border of the cell selection to move the selection, or drag the corner of the border to expand the selection.
- Press ENTER.

Create a Formula that Contains a Function: =AVERAGE(A1:B4)

The following formulas contain functions:

Example formula	What it does
=SUM(A:A)	Adds all numbers in column A
=AVERAGE(A1:B4)	Averages all numbers in the range

- Click the cell in which the formula enter has to be entered.
- To start the formula with the function, click “insert function” on the formula bar.
- Select the function.
- Enter the arguments. When the formula is completed, press ENTER.

Create a Formula with Nested Functions: =IF(AVERAGE(F2:F5)>50, SUM(G2:G5),0)
Nested functions use a function as one of the arguments of another function. The following formula sums a set of numbers (G2:G5) only if the average of another set of numbers (F2:F5) is greater than 50. Otherwise it returns 0.

Statistical Analysis Tools

Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that one can use to save steps when you develop complex statistical or engineering analyses. Provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Accessing the Data Analysis Tools: To access various tools included in the Analysis ToolPak click on “Data” menu, then click “Data Analysis” and select the appropriate analysis option. If the “Data Analysis” command is not available, we need to load the Analysis ToolPak “select and run the “Analysis ToolPack” from the “Add-Ins”.

Correlation

The “Correlation” analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. It can be used to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

If the experimenter had measured two variables in a group of individuals, such as foot-length and height, he/she can calculate how closely the variables are correlated with each other. Select “Data”, “Data Analysis”. Scroll down the list, select “Correlation” and click OK. A new window will appear where the following information needs to be entered:

Input range. Highlight the two columns of data that are the paired values for the two variables. The cell range will automatically appear in the box. If column headings are included in this range, tick the Labels box.

Output range. Click in this box then select a region on the worksheet where the user want the data table displayed. It can be done by clicking on a single cell, which will become the top left cell of the table.

Click OK and a table will be displayed showing the correlation coefficient (r) for the data.

CORREL(array1, array2) also returns the correlation coefficient between two data sets.

Covariance

Covariance is a measure of the relationship between two ranges of data. The “covariance” tool can be used to determine whether two ranges of data move together, *i.e.*, whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

To return the covariance for individual data point pairs, use the COVAR worksheet function.

Regression

The “Regression” analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables. For example, one can analyze how grain yield of barley is affected by factors like ears per plant, ear length (in cms), 100 grain weight (in gms) and number of grains per ear.

Descriptive Statistics

The “Descriptive Statistics” analysis tool generates a report of univariate statistics for data in the input range, which includes information about the central tendency and variability of the entered data.

Sampling

The “Sampling” analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, a representative sample can be used. One can also create a sample that contains only values from a particular part of a cycle if you believe that the input data is periodic. For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

Random Number Generation

The “Random Number Generation” analysis tool fills a range with independent random numbers drawn from one of several distributions. We can characterize subjects in a population with a probability distribution. For example, you might use a normal distribution to characterize the population of individuals' heights.

ANOVA: Single Factor

“ANOVA: Single Factor” option can be used for analysis of one-way classified data or data obtained from a completely randomized design. In this option, the data is given either in rows or columns such that observations in a row or column belong to one treatment only. Accordingly, define the input data range. Then specify whether, treatments are in rows or columns. Give the identification of upper most left corner cell in output range and click OK. In output, we get replication number of treatments, treatment totals, treatment means and treatment variances. In the ANOVA table besides usual sum of squares, Mean Square, F-calculated and P-value, it also gives the F-value at the pre-defined level of significance.

ANOVA: Two Factors with Replication

This option can be used for analysis of two-way classified data with m-observations per cell or for analysis of data obtained from a factorial CRD with two factors with same or different levels with same replications.

ANOVA: Two Factors without Replication

This option can be utilized for the analysis of two-way classified data with single observation per cell or the data obtained from a randomized complete block design. Suppose that there are ‘v’ treatments and ‘r’ replications and then prepare a v × r data sheet. Define it in input range, define alpha and output range.

t-Test: Two-Sample Assuming Equal Variances:

This analysis tool performs a two-sample student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal. TTEST(array1,array2,tails,type) returns the probability associated with a student's t test.

t-Test: Two-Sample Assuming Unequal Variances:

This t-test form assumes that the variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. Use this test when the groups under study are distinct.

t-Test: Paired Two Sample For Means:

This analysis tool performs a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. One can use this test when there is a natural pairing of observations in the samples, like a sample group is tested twice - before and after an experiment.

F-Test Two-Sample for Variances

The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances. For example, you can use an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams. FTEST(array1, array2) returns the result of an F-test, the one tailed probability that the variances of Array1 and array 2 are not significantly different.

Transformation of Data

The validity of analysis of variance depends on certain important assumptions like normality of errors and random effects, independence of errors, homoscedasticity of errors and effects are additive. The analysis is likely to lead to faulty conclusions when some of these assumptions are violated. A very common case of violation is the assumption regarding the constancy of variance of errors. One of the alternatives in such cases is to go for a weighted analysis of variance wherein each observation is weighted by the inverse of its variance. For this, an estimate of the variance of each observation is to be obtained which may not be feasible always. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Major aims of applying transformations are to bring data closer to normal distribution, to reduce relationship between mean and variance, to reduce the influence of outliers, to improve linearity in regression, to reduce interaction effects, to reduce skewness and kurtosis. Certain methods are available for identifying the transformation needed for any particular data set but one may also resort to certain standard forms of transformations depending on the nature of the data. Most commonly used transformations in the analysis of experimental data are Arcsine, Logarithmic and Square root. These transformations of data can be carried out using the following options.

Arcsine (ASIN): In the case of proportions, derived from frequency data, the observed proportion p can be changed to a new form $\theta = \sin^{-1(}\sqrt{p})$. This type of transformation is known as angular or arcsine transformation. However, when nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation. It may be noted that the angular transformation is not applicable to proportion or percentage data which are not derived from counts. For example, percentage of marks, percentage of profit, percentage of protein in grains, oil content in seeds, etc., can not be subjected to angular transformation. The angular transformation is not good when the data contain 0 or 1 values for p. The transformation in such cases is improved by replacing 0 with (1/4n) and 1 with [1-(1/4n)], before taking angular values, where *n* is the number of observations based on which p is estimated for each group.

ASIN gives the arcsine of a number. The arcsine is the angle whose sine is number and this number must be from -1 to 1. The returned angle is given in radians in the range $-\pi/2$ to $\pi/2$. To express the arcsine in degrees, multiply the result by 180/π. For this go to the CELL where the transformation is required and write =ASIN (Give Cell identification for which transformation to be done)* 180*7/22 and press ENTER. Then copy it for all observations.

Example: ASIN (0.5) equals 0.5236 (π/6 radians) and ASIN (0.5)* 180/PI equals 30 (degrees).

Logarithmic (LN): When the data are in whole numbers representing counts with a wide range, the variances of observations within each group are usually proportional to the squares of the group means. For data of this nature, logarithmic transformation is recommended. It squeezes the bigger values and stretches smaller values. A simple plot of group means against the group standard deviation will show linearity in such cases. A good example is data from an experiment involving various types of insecticides. For the effective insecticide, insect counts on the treated experimental unit may be small while for the ineffective ones, the counts may range from 100 to several thousands. When zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective in normalizing positively skewed distributions. It is also used to achieve additivity of effects in certain cases.

LN gives the natural logarithm of a positive number. Natural logarithms are based on the constant e (2.71828182845904). For this go the CELL where the transformation is required and write = LN(Give Cell Number for which transformation to be done) and press ENTER. Then copy it for all observations.

Example: LN(86) equals 4.454347, LN(2.7182818) equals 1, LN(EXP(3)) Equals 3 and EXP(LN(4)) equals 4. Further, EXP returns e raised to the power of a given number, LOG returns the logarithm of a number to a specified base and LOG 10 returns the base-10 logarithm of a number.

Square Root (SQRT): If the original observations are brought to square root scale by taking the square root of each observation, it is known as square root transformation. This is appropriate when the variance is proportional to the mean as discernible from a graph of group variances against group means. Linear relationship between mean and variance is commonly observed when the data are in the form of small whole numbers (*e.g.*, counts of wildlings per quadrat, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.). When the observed values fall within the range of 1 to 10 and especially when zeros are present, the transformation should be, $\sqrt{}(y + 0.5)$.

SQRT gives square root of a positive number. For this go to the CELL where the transformation is required and write = SQRT (Give Cell No. for which transformation to be done = 0.5) and press ENTER. Then copy it for all observations. However, if number is negative, SQRT return the #NUM ! error value.

Example: SQRT(16) equals 4, SQRT(-16) equals #NUM! and SQRT(ABS(-16)) equals 4.

Once the transformation has been made, the analysis is carried out with the transformed data and all the conclusions are drawn in the transformed scale. However, while presenting the results, the means and their standard errors are transformed back into original units. While transforming back into the original units, certain corrections have to be made for the means. In the case of log transformed data, if the mean value is $\bar{\mathbf{y}}$, the mean value of the original units will be antilog ($\bar{\mathbf{y}}$ + 1.15$\bar{\mathbf{y}}$) instead of antilog ($\bar{\mathbf{y}}$). If the square root transformation had been used, then the mean in the original scale would be antilog $((\bar{\mathbf{y}} + V(\bar{\mathbf{y}}))^2$ instead of $(\bar{\mathbf{y}})^2$ where $V(\bar{\mathbf{y}})$ represents the variance of $\bar{\mathbf{y}}$. No such correction is generally made in the case of angular transformation. The inverse transformation for angular transformation would be $p = (\sin q)^2$.

Sum(SUM): It gives the sum of all the numbers in the list of arguments. For this go to the CELL where the sum of observations is required and write = SUM (define data range for which the sum is required) and press ENTER. Instead of defining the data range, the exact numerical values to be added can also be given in the argument viz. SUM (Number1, number2,…), number1, number2,… are 1 to 30 arguments for which you want the sum.

Example: If cells A2:E2 contain 5, 15,30,40 and 50; SUM(A2:C2) equals 50, SUM(B2:E2,15) equals 150 and SUM(5,15) equals 20.

Some other related functions with this option are:

AVERAGE returns the average of its arguments, PRODUCT multiplies its arguments and SUMPRODUCT returns the sum of the products of corresponding array components.

Sum of Squares (SUMSQ): This gives the sum of the squares of the list of arguments. For this go to the CELL where the sum of squares of observations is required and write = SUMSQ (define data range for which the sum of squares is required) and press ENTER.

Example: If cells A2:E2 contain 5, 15, 30, 40 and 50; SUMSQ(A2:C2) equals 1150 and SUMSQ(3,4) equals 25.

Matrix Multiplication (MMULT): It gives the matrix product of two arrays, say array 1 and array 2. The result is an array with the same number of rows as array1, say a and the same number of columns as array2, say b. For getting this mark the a × b cells on the spread sheet. Write =MMULT (array 1, array 2) and press Control +Shift+ Enter. The number of columns in array1 must be the same as the number of rows in array2, and both arrays must contain only numbers. Array1 and array2 can be given as cell ranges, array constants, or references. If any cells are empty or contain text, or if the number of columns in array1 is different from the number of rows in array2, MMULT returns the ≠VALUE! error value.

Determinant of a Matrix (MDETERM): It gives the value of the determinant associated with the matrix. Write = MDETERM(array) and press Control + Shift + Enter.

Matrix Inverse (MINVERSE): It gives the inverse matrix for the non-singular matrix stored in a square array, say of order p. i.e., an array with equal number of rows and columns. For getting this mark the p × p cells on the spread sheet where the inverse of the array is required and write = MINVERSE(array) and press Control + Shift + Enter. Array can be given as a cell range, such as A1:C3; as an array constant, such as {1,2,3;4,5,6;7,8,8}; or as a name for either of these. If any cells in array are empty or contain text, MINVERSE returns the ≠VALUE! error value.

Example: MINVERSE ({4,-1;2,0}) equals {0,0.5;-1,2}and MINVERSE ({1,2,1;3,4,-1;0,2,0}) equals {0.25, 0.25,-0.75;0,0,0.5;0.75,-0.25,-0.25}.

Transpose (TRANSPOSE): For getting the transpose of an array mark the array and then select copy from the EDIT menu. Go to the left corner of the array where the transpose is required. Select the EDIT menu and then paste special and under paste special select the TRANSPOSE option.

Exercises on MS-Excel

1. Table below contains values of pH and organic carbon content observed in soil samples collected from natural forest. Compute mean, median, standard deviation, range and skewness of the data.

Soil pit	**pH (x)**	**Organic carbon (%) (y)**		**Soil pit**	**pH (x)**	**Organic carbon (%) (y)**
1	5.7	2.10		9	5.4	2.09
2	6.1	2.17		10	5.9	1.01
3	5.2	1.97		11	5.3	0.89
4	5.7	1.39		12	5.4	1.60
5	5.6	2.26		13	5.1	0.90
6	5.1	1.29		14	5.1	1.01
7	5.8	1.17		15	5.2	1.21
8	5.5	1.14				

2. Consider the following data on various characteristics of a crop:

pp	**ph**	**ngl**	**yield**
142	0.525	8.2	2.47
143	0.64	9.5	4.76
107	0.66	9.3	3.31
78	0.66	7.5	1.97
100	0.46	5.9	1.34
86.5	0.345	6.4	1.14
103.5	0.86	6.4	1.5
155.99	0.33	7.5	2.03
80.88	0.285	8.4	2.54
109.77	0.59	10.6	4.9
61.77	0.265	8.3	2.91
79.11	0.66	11.6	2.76
155.99	0.42	8.1	0.59
61.81	0.34	9.4	0.84
74.5	0.63	8.4	3.87
97	0.705	7.2	4.47
93.14	0.68	6.4	3.31
37.43	0.665	8.4	1.57
36.44	0.275	7.4	0.53
51	0.28	7.4	1.15
104	0.28	9.8	1.08

49	0.49	4.8	1.83
54.66	0.385	5.5	0.76
55.55	0.265	5	0.43
88.44	0.98	5	4.08
99.55	0.645	9.6	2.83
63.99	0.635	5.6	2.57
101.77	0.29	8.2	7.42
138.66	0.72	9.9	2.62
90.22	0.63	8.4	2

(i) Sort yield in ascending order and filter the data ph less than 0.3 or greater than 0.6 from the data.

(ii) Find the correlation coefficient and fit the multiple regression equation by taking yield as dependent variable.

3. Let **A**, **B** and **C** be three matrices as follows:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 6 & 1 & 9 \\ 3 & 5 & 6 & 7 & 2 \\ 8 & 3 & 9 & 1 & 5 \\ 3 & 1 & 1 & 1 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 2 & 4 \\ 1 & 9 \\ 8 & 1 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 2 & 3 & 1 & 8 & 4 \\ 3 & 6 & 7 & 8 & 8 \\ 2 & 3 & 5 & 5 & 7 \\ 2 & 3 & 6 & 6 & 1 \\ 1 & 2 & 8 & 5 & 5 \end{bmatrix}.$$

Find (i) **AB** (ii) $\mathbf{C}^{-1}$ (iii) $|\mathbf{A}|$ (iv) $\mathbf{A}^{T}$.

4. Draw line graph for the following data on a tree species:

Year	Height (cm)	Diameter
1981	21	5.0
1982	34	8.0
1983	11	9.0
1984	13	3.0
1985	15	2.4
1986	55	5.5
1987	30	6.9
1988	50	9.1
1989	23	10.0
1990	22	2.5
1991	37	3.4
1992	38	6.2
1993	37	7.0
1994	11	8.1
1995	20	9.0
1996	16	3.7
1997	54	9.0
1998	33	4.0
1999	12	6.7
2000	19	7.7

Also draw a bar diagram using the above data.

5. The table below lists plant height in cm of seedlings of rice belonging to the two varieties. Examine whether the two samples are coming from populations having equal variance, using F-test. Further, test whether the average height of the two groups are the same, using appropriate t-test.

Plot	Group I	Group II
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2
7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

6. Examine whether the average organic carbon content measured from two layers of a set of soil pits from a pasture are same using paired t-test from the data given below:

	Organic carbon (%)	
Soil pit	Layer 1 (x)	Layer 2 (y)
1	1.59	1.21
2	1.39	0.92
3	1.64	1.31
4	1.17	1.52
5	1.27	1.62
6	1.58	0.91
7	1.64	1.23
8	1.53	1.21
9	1.21	1.58
10	1.48	1.18

7. Mycelial growth in terms of diameter of the colony (mm) of *R. solani* isolates on PDA medium after 14 hours of incubation is given in the table below. Carry out the CRD analysis for the data. And draw your inferences.

R. solani isolates	Mycelial growth		
	Repl. 1	Repl. 2	Repl. 3
RS 1	29.0	28.0	29.0
RS 2	33.5	31.5	29.0
RS 3	26.5	30.0	
RS 4	48.5	46.5	49.0
RS 5	34.5	31.0	

8. Following is the data on mean yield in kg per plot of an experiment conducted to compare the performance of 8 treatments using a Randomized Complete Block design with 3 replications. Perform the analysis of variance.

Treatment (Provenance)	**Replication**		
	I	**II**	**III**
1	30.85	38.01	35.10
2	30.24	28.43	35.93
3	30.94	31.64	34.95
4	29.89	29.12	36.75
5	21.52	24.07	20.76
6	25.38	32.14	32.19
7	22.89	19.66	26.92
8	29.44	24.95	37.99

9. From the following data make a summary table for finding out the average of X_9 for various years and various levels of X_6 using pivot table and pivot chart report option of MS-Excel.

YR	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1995	1	1	40	30	0	60	40	4861	5208	5556	5694
1995	1	2	40	30	0	60	40	4167	4444	4861	5035
1995	2	3	40	30	0	60	40	4618	4653	4653	5174
1995	2	4	40	30	0	60	40	4028	4167	4514	4722
1995	2	5	40	30	0	60	40	4306	4514	4653	4861
1996	2	1	40	30	0	60	40	6000	5750	5499	6250
1996	2	2	40	30	0	60	40	5646	5000	5250	5444
1996	2	3	40	30	0	60	40	4799	5097	4896	5299
1996	2	4	40	30	0	60	40	5250	5299	4194	4847
1996	3	1	40	30	0	60	40	5139	5417	5764	5903
1996	3	2	40	30	0	60	40	5417	5694	6007	6111
1996	4	1	40	30	0	60	40	6300	7450	7750	8000
1996	4	2	40	30	0	60	40	6350	7850	7988	8200
1996	4	3	40	30	0	60	40	5750	6400	6600	6700
1996	4	4	40	30	0	60	40	6000	7250	7450	7681
1996	5	1	40	30	0	60	40	3396	4090	5056	5403
1996	5	2	40	30	0	60	40	5194	5000	6000	6500
1996	5	3	40	30	0	60	40	4299	4250	4750	5250
1996	6	1	40	30	0	60	40	4944	5194	5000	5097
1996	6	2	40	30	0	60	40	5395	5499	5499	5597
1996	6	3	40	30	0	60	40	3444	5646	5000	5000
1996	6	4	40	30	0	60	40	6250	6500	6646	6750
1997	1	1	120	30	30	120	60	5839	6248	6199	6335
1997	1	2	120	30	30	120	60	5590	5652	5702	5851
1997	2	1	120	30	30	120	60	4497	4794	4894	5205
1997	2	2	120	30	30	120	60	4696	5006	5304	5702
1997	2	3	120	30	30	120	60	4398	4596	4894	5304
1997	2	4	120	30	30	120	60	4497	5503	5702	6099

1997	3	1	120	30	30	120	60	4199	5602	5801	6000
1997	3	2	120	30	30	120	60	3404	3901	4199	4497
1997	3	3	120	30	30	120	60	3602	5404	5503	5801
1997	3	4	120	30	30	120	60	3602	4297	4497	4696
1997	4	1	120	30	30	120	60	3205	3801	4199	4894
1997	4	2	120	30	30	120	60	3801	4794	6099	6298
1997	4	3	120	30	30	120	60	3503	5205	6298	6795
1997	4	4	120	30	30	120	60	3205	4894	5503	6199
1997	5	1	120	30	30	120	60	4199	4099	4199	4297
1997	5	2	120	30	30	120	60	3304	3702	3602	3801
1997	5	3	120	30	30	120	60	2596	2894	3106	3205
1998	1	1	40	30	0	60	40	3727	3106	3404	3503
1998	1	2	40	30	0	60	40	4894	4348	4447	4534
1998	1	3	40	30	0	60	40	2696	2795	3056	3205
1998	2	2	40	30	0	60	40	5503	4298	4497	4795
1998	2	3	40	30	0	60	40	5006	3702	3702	3901

10. From the data given in problem 10, sort X_{10} in ascending order. Also, filter the data for $X_{11} < 4200$ or $X_{11} > 5000$.

R SOFTWARE: AN OVERVIEW

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R is a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

R environment

The R environment provides an integrated suite of software facilities for data manipulation, calculation and graphical display. It has

- a data handling and storage facility,
- a suite of operators for calculations on arrays and matrices,
- a large, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display, and
- a well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined functions and input and output facilities.

Origin

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland started the project on R in 1995 and hence the name software has been named as 'R'.

R was introduced as an environment within which many classical and modern statistical techniques can be implemented. A few of these are built into the base R environment, but many are supplied as packages. There are a number of packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via http://cran.r-project.org) and elsewhere.

Availability

Since R is an open source project, it can be obtained freely from the website https://www.r-project.org/. One can download R from any CRAN mirror out of several CRAN (Comprehensive R Archive Network) mirrors. Latest available version of R is *R version 3.6.0* and it has been released on 26.04.2019.

Installation

To install R in windows operating system, simply double click on the setup file. It will automatically install the software in the system.

Usage

R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windows set up only.

Difference with other packages

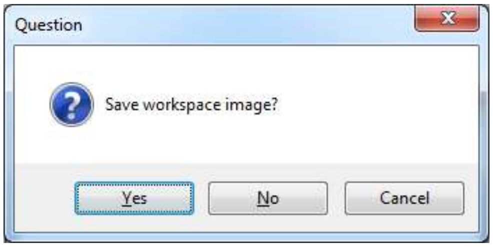
There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects.

Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output and store the results in an object for subsequent interrogation by further R functions.

Invoking R

If properly installed, usually R has a shortcut icon on the desktop screen and/or you can find it under Start|All Programs|R menu.

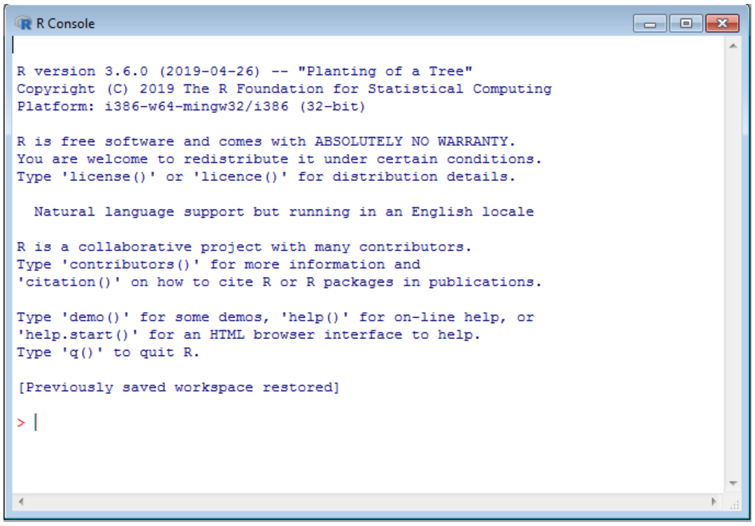
To quit R, type *q()* at the R prompt (>) and press Enter key. A dialog box will ask whether to save the objects you have created during the session so that they will become available next time when R will be invoked.



Save workspace image?

Windows of R

R has only one window and when R is started it looks like



```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

R commands

i. R commands are case sensitive, so X and x are different symbols and would refer to different variables.
ii. Elementary commands consist of either expressions or assignments.
iii. If an expression is given as a command, it is evaluated, printed and the value is lost.

iv. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed.
v. Commands are separated either by a semi-colon (';'), or by a newline.
vi. Elementary commands can be grouped together into one compound expression by braces '{' and '}'.
vii. Comments can be put almost anywhere, starting with a hashmark ('#'). Anything written after # marks to the end of the line is considered as a comment.
viii. Window can be cleared of lines by pressing Ctrl + L keys.

Executing commands from or diverting output to a file

If commands are stored in an external file, say 'D:/commands.txt' they may be executed at any time in an R session with the command

```
> source("d:/commands.txt")
```

For Windows Source is also available on the File menu.

The function *sink()*,

```
> sink("d:/record.txt")
```

will divert all subsequent output from the console to an external file, 'record.txt' in D drive. The command

```
> sink()
```

restores it to the console once again.

Simple manipulations of numbers and vectors

R operates on named data structures. The simplest such structure is the numeric vector, which is a single entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

The function *c()* assigns the five numbers to the vector x. The assignment operator (<-) 'points' to the object receiving the value of the expression. Once can use the '=' operator as an alternative.

A single number is taken as a vector of length one.

Assignments can also be made in the other direction, using the obvious change in the assignment operator. So the same assignment could be made using

```
> c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed. So now if we were to use the command

```
> 1/x
```

the reciprocals of the five values would be printed at the terminal.

The elementary arithmetic operators
+ addition
– subtraction
* multiplication
/ division
^ exponentiation

Arithmetic functions
log, exp, sin, cos, tan, sqrt,

Other basic functions
max(x) – maximum element of vector x,
min(x)- minimum element of vector x,
range (x) – range of the values of vector x ,
length(x) - the number of elements in x,
sum(x) - the total of the elements in x,
prod(x) – product of the elements in x
mean(x) – average of the elements of x
var(x) – sample variance of the elements of (x)
sort(x) – returns a vector with elements sorted in increasing order.

Logical operators
< - less than
<= less than or equal to
> greater than
>= greater than or equal to
== equal to
!= not equal to.

Other objects in R
Matrices or arrays - multi-dimensional generalizations of vectors.
Lists - a general form of vector in which the various elements need not be of the same type, and are often themselves vectors or lists.
Functions - objects in R which can be stored in the project's workspace. This provides a simple and convenient way to extend R.

Matrix facilities
A matrix is just an array with two subscripts. R provides many operators and functions those are available only for matrices. Some of the important R functions for matrices are
t(A) – transpose of the matrix A
nrow(A) – number of rows in the matrix A
ncol(A) – number of columns in the matrix A
A%*% B– Cross product of two matrices A and B
A*B – element by element product of two matrices A and B
diag (A) – gives a vector of diagonal elements of the square matrix A
diag(a) – gives a matrix with diagonal elements as the elements of vector a

eigen(A) – gives eigen values and eigen vectors of a symmetric matrix A
rbind (A,B) – concatenates two matrix A and B by appending B matrix below A matrix (They should have same number of columns)
cbind(A, B) - concatenates two matrix A and B by appending B matrix in the right of A matrix (They should have same number of rows)

Data frame
Data frame is an array consisting of columns of various mode (numeric, character, etc). Small to moderate size data frame can be constructed by *data.frame()* function. For example, following is an illustration how to construct a data frame from the car data*:

Make	**Model**	**Cylinder**	**Weight**	**Mileage**	**Type**
Honda	Civic	V4	2170	33	Sporty
Chevrolet	Beretta	V4	2655	26	Compact
Ford	Escort	V4	2345	33	Small
Eagle	Summit	V4	2560	33	Small
Volkswagen	Jetta	V4	2330	26	Small
Buick	Le Sabre	V6	3325	23	Large
Mitsubishi	Galant	V4	2745	25	Compact
Dodge	Grand Caravan	V6	3735	18	Van
Chrysler	New Yorker	V6	3450	22	Medium
Acura	Legend	V6	3265	20	Medium

```
> Make<-c("Honda","Chevrolet","Ford","Eagle","Volkswagen","Buick","Mitsbusihi",
+ "Dodge","Chrysler","Acura")
> Model=c("Civic","Beretta","Escort","Summit","Jetta","Le Sabre","Galant",
+ "Grand Caravan","New Yorker","Legend")
```

Note that the plus sign (+) in the above commands are automatically inserted when the carriage return is pressed without completing the list. Save some typing by using *rep()* command. For example, *rep("V4",5)* instructs R to repeat V4 five times.

```
> Cylinder<-c(rep("V4",5),"V6","V4",rep("V6",3))
> Cylinder
 [1] "V4" "V4" "V4" "V4" "V4" "V6" "V4" "V6" "V6" "V6"
> Weight<-c(2170,2655,2345,2560,2330,3325,2745,3735,3450,3265)
> Mileage<-c(33,26,33,33,26,23,25,18,22,20)
> Type<-c("Sporty","Compact",rep("Small",3),"Large","Compact","Van",rep("Medium",2))
```

Now *data.frame()* function combines the six vectors into a single data frame.

```
> Car<-data.frame(Make,Model,Cylinder,Weight,Mileage,Type)
> Car
        Make         Model Cylinder Weight Mileage    Type
1      Honda         Civic       V4   2170      33  Sporty
2  Chevrolet       Beretta       V4   2655      26 Compact
3       Ford        Escort       V4   2345      33   Small
```

```
4       Eagle        Summit       V4   2560       33   Small
5  Volkswagen         Jetta       V4   2330       26   Small
6       Buick      Le Sabre       V6   3325       23   Large
7  Mitsbusihi        Galant       V4   2745       25 Compact
8       Dodge Grand Caravan       V6   3735       18     Van
9    Chrysler    New Yorker       V6   3450       22  Medium
10      Acura        Legend       V6   3265       20  Medium

> names(Car)
[1] "Make"     "Model"    "Cylinder" "Weight"   "Mileage"  "Type"
```

Just as in matrix objects, partial information can be easily extracted from the data frame:

```
> Car[1,]

   Make Model Cylinder Weight Mileage   Type

1 Honda Civic       V4   2170      33 Sporty
```

In addition, individual columns can be referenced by their labels:

```
> Car$Mileage
 [1] 33 26 33 33 26 23 25 18 22 20
> Car[,5]        #equivalent expression
> mean(Car$Mileage)    #average mileage of the 10 vehicles
[1] 25.9
> min(Car$Weight)
[1] 2170
```

table() command gives a frequency table:

```
> table(Car$Type)

Compact   Large  Medium   Small  Sporty     Van
      2       1       2       3       1       1
```

If the proportion is desired, type the following command instead:

```
> table(Car$Type)/10

Compact   Large  Medium   Small  Sporty     Van
    0.2     0.1     0.2     0.3     0.1     0.1
```

Note that the values were divided by 10 because there are that many vehicles in total. If you don't want to count them each time, the following does the trick:

```
> table(Car$Type)/length(Car$Type)
```

Cross tabulation is very easy, too:

```
> table(Car$Make, Car$Type)
```

```
              Compact Large Medium Small Sporty Van
  Acura       0       0     1      0     0      0
  Buick       0       1     0      0     0      0
  Chevrolet   1       0     0      0     0      0
  Chrysler    0       0     1      0     0      0
  Dodge       0       0     0      0     0      1
  Eagle       0       0     0      1     0      0
  Ford        0       0     0      1     0      0
  Honda       0       0     0      0     1      0
  Mitsbusihi  1       0     0      0     0      0
  Volkswagen  0       0     0      1     0      0
```

What if you want to arrange the data set by vehicle weight? *order()* gets the job done.

```
> i<-order(Car$Weight);i
 [1] 1 5 3 4 2 7 10 6 9 8
> Car[i,]
         Make         Model Cylinder Weight Mileage    Type
1       Honda         Civic       V4   2170      33  Sporty
5  Volkswagen         Jetta       V4   2330      26   Small
3        Ford        Escort       V4   2345      33   Small
4       Eagle        Summit       V4   2560      33   Small
2   Chevrolet       Beretta       V4   2655      26 Compact
7  Mitsbusihi        Galant       V4   2745      25 Compact
10      Acura        Legend       V6   3265      20  Medium
6       Buick      Le Sabre       V6   3325      23   Large
9    Chrysler    New Yorker       V6   3450      22  Medium
8       Dodge Grand Caravan       V6   3735      18     Van
```

Creating/editing data objects

```
> y<-c(1,2,3,4,5);y
[1] 1 2 3 4 5
```

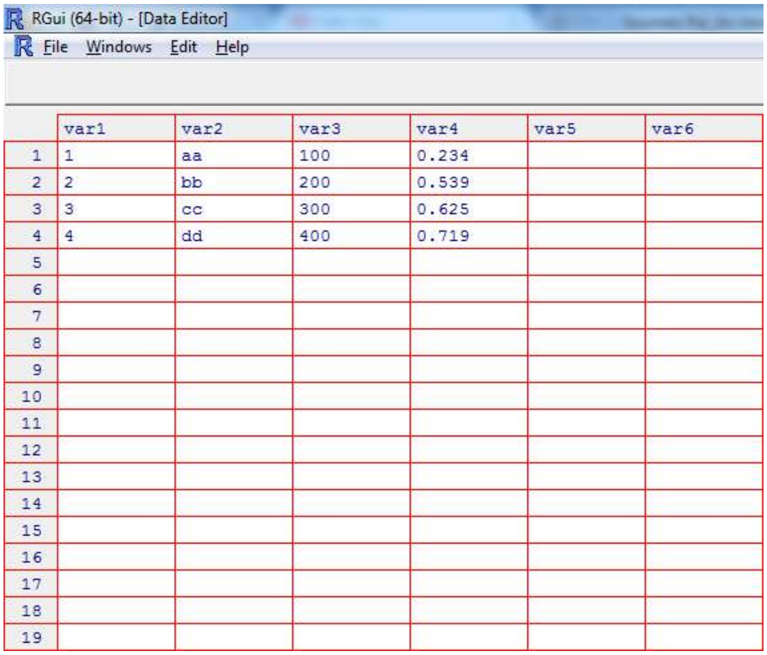
If you want to modify the data object, use *edit()* function and assign it to an object. For example, the following command opens R Editor for editing.

```
> y<-edit(y)
```

If you prefer entering the data.frame in a spreadsheet style data editor, the following command invokes the built-in editor with an empty spreadsheet.

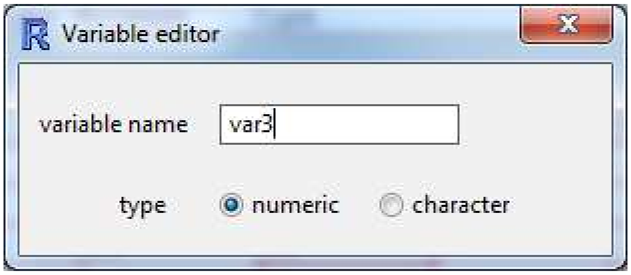
```
> data1<-edit(data.frame())
```

After entering a few data points, it looks like this:



	var1	var2	var3	var4	var5	var6
1	1	aa	100	0.234		
2	2	bb	200	0.539		
3	3	cc	300	0.625		
4	4	dd	400	0.719		
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

You can also change the variable name by clicking once on the cell containing it. Doing so opens a dialog box:



variable name var3

type ● numeric ○ character

When finished, click ✕ in the upper right corner of the dialog box to return to the Data Editor window. Close the Data Editor to return to the R command window (R Console). Check the result by typing:

```
> data1
```

Reading data from files

When data files are large, it is better to read data from external files rather than entering data through the keyboard. To read data from an external file directly, the external file should be arranged properly.

The first line of the file should have a name for each variable. Each additional line of the file has the values for each variable.

Input file form with names and row labels:

Price	Floor	Area	Rooms	Age	isNew
52.00	111.0	830	5	6.2	no

54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	yes
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes

...

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as isNew in the example, as factors. This can be changed if necessary.

The function *read.table()* can then be used to read the data frame directly

```
> HousePrice <- read.table("d:/houses.data", header = TRUE)
```

Reading comma delimited data

The following commands can be used for reading comma delimited data into R.

read.csv(filename) This command reads a .CSV file into R. You need to specify the exact filename with path.

read.csv(file.choose()) This command reads a .CSV file but the *file.choose()* part opens up an explorer type window that allows you to select a file from your computer. By default, R will take the first row as the variable names.

read.csv(file.choose(), header=T)

This reads a .CSV file, allowing you to select the file, the header is set explicitly. If you change to header=F then the first row will be treated like the rest of the data and not as a label.

Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,

```
>attach(HousePrice)
```

causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice to use the *attach(datafile)* function immediately after reading the *datafile* into R.

Packages

All R functions and datasets are stored in packages. The contents of a package are available only when the package is loaded. This is done to run the codes efficiently without much memory usage. To see which packages are installed at your machine, use the command

```
> library()
```

To load a particular package, use a command like

```
> library(forecast)
```

Users connected to the Internet can use the *install.packages()* and *update.packages()* functions to install and update packages. Use *search()* to display the list of packages that are loaded.

Standard packages

The standard (or base) packages are considered part of the R source code. They contain the basic functions those allow R to work with the datasets and standard statistical and graphical functions. They should be automatically available in any R installation.

Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (https://cran.r-project.org/web/packages/), and other repositories such as Bioconductor (http://www.bioconductor.org/). The collection of available packages changes frequently. As on June 07, 2019, the CRAN package repository contains 14346 available packages.

Getting Help

Complete help files in HTML and PDF forms are available in R. To get help on a particular command/function etc., type *help (command name)*. For example, to get help on function 'mean', type *help(mean)* as shown below

```
> help(mean)
```

This will open the help file with the page containing the description of the function mean.
Another way to get help is to use "?" followed by function name. For example,

```
>?mean
```

will open the same window again.

In this lecture note, all R commands and corresponding outputs are given in `Courier New` font to differentiate from the normal texts. Since R is case-sensitive, i.e. typing *Help(mean),* would generate an error message,

```
> Help(mean)

Error in Help(mean) : could not find function "Help"
```

Further Readings

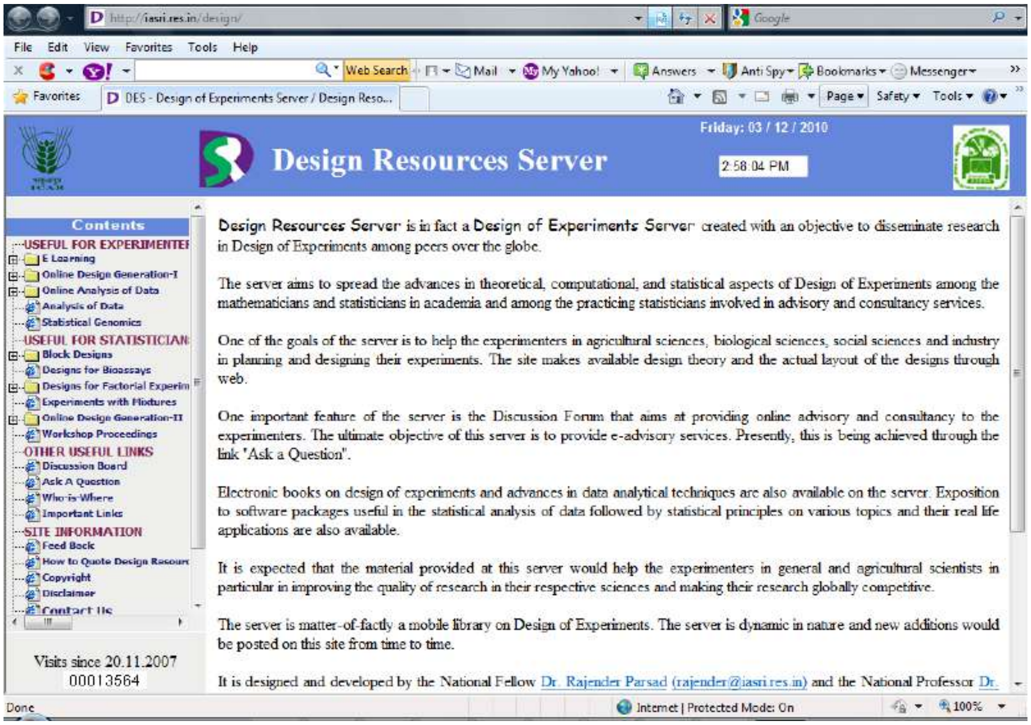
Various documents are available in https://cran.r-project.org/manuals.html from beginners' level to most advanced level. The following manuals are available in pdf form:

1. An Introduction to R
2. R Data Import/Export
3. R Installation and Administration
4. Writing R Extensions
5. The R language definition
6. R Internals
7. The R Reference Index

DESIGN RESOURCES SERVER

1. Introduction

Design Resources Server is developed to popularize and disseminate the research in Design of Experiments among the scientists of National Agricultural Research System (NARS) in particular and researchers all over the globe in general and is hosted at www.iasri.res.in/design. The home page of the server is



Design Resources Server is in fact a Design of Experiments Server created with an objective to disseminate research in Design of Experiments among peers over the globe.

The server aims to spread the advances in theoretical, computational, and statistical aspects of Design of Experiments among the mathematicians and statisticians in academia and among the practicing statisticians involved in advisory and consultancy services.

One of the goals of the server is to help the experimenters in agricultural sciences, biological sciences, social sciences and industry in planning and designing their experiments. The site makes available design theory and the actual layout of the designs through web.

One important feature of the server is the Discussion Forum that aims at providing online advisory and consultancy to the experimenters. The ultimate objective of this server is to provide e-advisory services. Presently, this is being achieved through the link 'Ask a Question".

Electronic books on design of experiments and advances in data analytical techniques are also available on the server. Exposition to software packages useful in the statistical analysis of data followed by statistical principles on various topics and their real life applications are also available.

It is expected that the material provided at this server would help the experimenters in general and agricultural scientists in particular in improving the quality of research in their respective sciences and making their research globally competitive.

The server is matter-of-factly a mobile library on Design of Experiments. The server is dynamic in nature and new additions would be posted on this site from time to time.

It is designed and developed by the National Fellow Dr. Rajender Parsad (rajender@iasri.res.in) and the National Professor Dr.

Design Resources Server is matter-of-factly a virtual, mobile library on design of experiments created with an objective to advise and help the experimenters in agricultural sciences, biological sciences, animal sciences, social sciences and industry in planning and designing their experiments for making precise and valid inferences on the problems of their interest. This also provides support for analysis of data generated so as to meet the objectives of the study. The server also aims at providing a platform to the researchers in design of experiments for disseminating research and also strengthening research in newer emerging areas so as to meet the challenges of agricultural research. The purpose of this server is to spread advances in theoretical, computational, and statistical aspects of Design of Experiments among the mathematicians and statisticians in academia and among the practicing statisticians involved in advisory and consultancy services.

This server works as an e-advisory resource for the experimenters. The actual layout of the designs is available to the experimenters online and the experimenter can use these designs for their experimentation. It is expected that the material provided at this server would help the experimenters in general and agricultural scientists in particular in improving the quality of research in their respective sciences and making their research globally competitive.

Design Server is open to everyone from all over the globe. Anyone can join this and add information to the site to strengthen it further with the permission of the developers. The Server contains a lot of useful information for scientists of NARS. The material available on the server has been partitioned into 4 components:

- **Useful for Experimenters:** Electronic Books, online generation of randomized layout of designs, online analysis of data, analysis of data using various softwares, statistical genomics.
- **Useful for Statisticians:** Literature and catalogues of BBB designs, designs for making test treatments-control treatment comparisons, designs for bioassays, designs for factorial experiments (supersaturated designs, block designs with factorial treatment structure), experiments with mixtures, Online generation of Hadamard matrices, MOLS and orthogonal arrays.
- **Other Useful Links:** Discussion Board, Ask a Question, Who-is-where, important links.
- **Site Information:** Feedback, How to Quote Design Resources Server, Copyright, disclaimer, contact us and site map.

The major components are Useful for Experimenters and Research Statisticians. The scientists, however, can use either of the parts or parts of their choice. A brief description of all the above four components is given in the sequel.

2. Useful for Experimenters

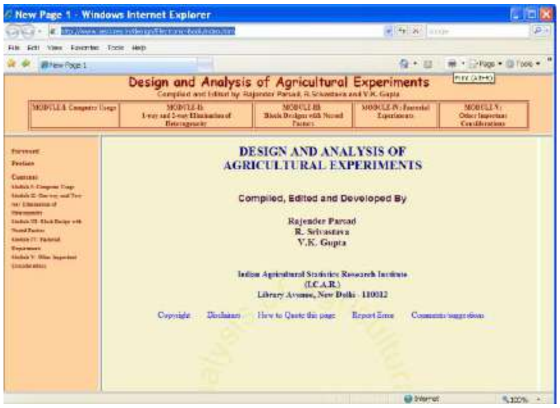
This link has been designed essentially to meet the requirements of the experimenters whose prime interest is in designing the experiment and then subsequently analyzing the data generated so as to draw statistically valid inferences. To meet this end, the link contains the following sub-links:

2.1 E-Learning

This is an important link that provides useful and important reading material on use of some statistical software packages, designing experiments, statistical analysis of data and other useful topics in statistics in the form of two electronic books viz.

1. Design and Analysis of Agricultural Experiments
   www.iasri.res.in/design/Electronic-Book/index.htm
2. Advances in Data Analytical Techniques
   www.iasri.res.in/design/ebook/EBADAT/index.htm

The screen shots of cover pages of these books are shown below:


New Page 1 - Windows Internet Explorer
File Edit View Favorites Tools Help
New Page 1
Page
Tools
Internet
100%

Design and Analysis of Agricultural Experiments

Compiled and Edited by Rajender Parsad, R. Srivastava and V.K. Gupta

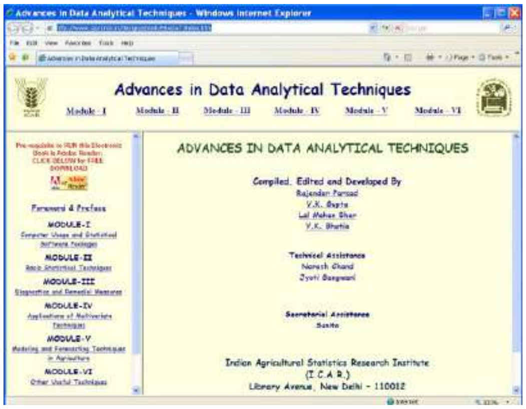
DESIGN AND ANALYSIS OF AGRICULTURAL EXPERIMENTS

Compiled, Edited and Developed By

Rajender Parsad
R. Srivastava
V.K. Gupta

Indian Agricultural Statistics Research Institute
(I.C.A.R.)
Library Avenue, New Delhi - 110012

Copyright Disclaimer How to Quote this page Report Error Comments/suggestions


Advances in Data Analytical Techniques - Windows Internet Explorer
File Edit View Favorites Tools Help
Advances in Data Analytical Techniques
Page
Tools
Internet

Advances in Data Analytical Techniques

Module - I Module - II Module - III Module - IV Module - V Module - VI

Foreword & Preface
MODULE-I
Computer Usage and Statistical Software Packages
MODULE-II
Basic Statistical Techniques
MODULE-III
Diagnostics and Remedial Measures
MODULE-IV
Applications of Multivariate Techniques
MODULE-V
Modeling and Forecasting Techniques in Agriculture
MODULE-VI
Other Useful Techniques

ADVANCES IN DATA ANALYTICAL TECHNIQUES

Compiled, Edited and Developed By
Rajender Parsad
V.K. Gupta
Lal Mohan Bhar
V.K. Bhatia

Technical Assistance
Naresh Chand
Jyoti Bangwani

Secretarial Assistance
Sunita

Indian Agricultural Statistics Research Institute
(I.C.A.R.)
Library Avenue, New Delhi - 110012

The coverage of topics in these electronic books is very wide and almost all the aspects of designing an experiment and analysis of data are covered. The chapters are decorated with solved examples giving the steps of analysis. The users can have online access to these electronic books. This provides good theoretical support and also reading material to the users.

2.2 Online Design Generation-I

This link is very useful for experimenters because it helps in generation of randomized layout of the following designs:

Basic Designs: Generates of randomized layout of completely randomized design and randomized complete block design both for single factor and multifactor experiments and Latin square designs for single factor experiments. The field book can be created as a .csv file or a text file. This is available at

www.iasri.res.in/design/Basic Designs/generate_designs.htm.

Augmented Designs: A large number of germplasm evaluation trials are conducted using augmented designs. The experimenters generally compromise with the randomization of treatments in the design. Further, experimenters also need to know the optimum replication number of controls in each block so as to maximize the efficiency per observation. Online software for generation of randomized layout of an augmented randomized complete block design for given number of test treatments, control treatments and number of blocks with given block sizes, not necessarily equal, is developed and is available at

www.iasri.res.in/design/Augmented Designs/home.htm.

The design can be generated with optimum replication of control treatments in each block so as to maximize efficiency per observation.

Resolvable Block Designs: Resolvable block designs are an important class of incomplete block designs wherein the blocks can be formed together into sets with the blocks within each set constituting a complete replication. In the class of resolvable block designs, square lattice designs are very popular among experimenters. One can generate square lattice designs with three replications using

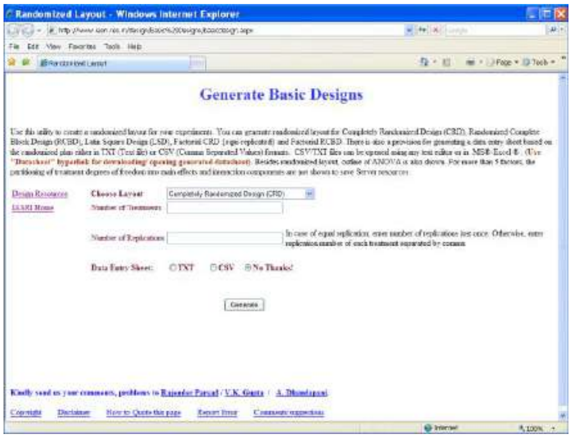
www.iasri.res.in/WebHadamard/square lattice.htm.

Another important class of resolvable block designs is the alpha designs. These designs are available when the number of treatments is a composite number. Literature on alpha designs is available at

www.iasri.res.in/design/Alpha/Home.htm.

This link also provides randomized layout of alpha designs for $6 \leq v$ (=sk, the number of treatments) ≤ 150, $2 \leq r$ (number of replications) ≤ 5, $3\leq k$ (block size) ≤ 10 and $2 \leq s \leq 15$ along with the lower bounds to A- and D- efficiencies of the designs.

The screen shots for generation of randomized layout of basic designs, augmented designs, square lattice designs and alpha designs are



Randomized Layout - Windows Internet Explorer
File Edit View Favorites Tools Help

Generate Basic Designs

Use this utility to create a randomized layout for your experiments. You can generate randomized layout for Completely Randomized Design (CRD), Randomized Complete Block Design (RCBD), Latin Square Design (LSD), Factorial CRD (equi-replicated) and Factorial RCBD. There is also a provision for generating a data entry sheet based on the randomized plan either in TXT (Text file) or CSV (Comma Separated Values) formats. CSV/TXT files can be opened using any text editor or in MS® Excel ®. (Use "Datasheet" hyperlink for downloading/ opening generated datasheet). Besides randomized layout, outline of ANOVA is also shown. For more than 5 factors, the partitioning of treatment degrees of freedom into main effects and interaction components are not shown to save Server resources.

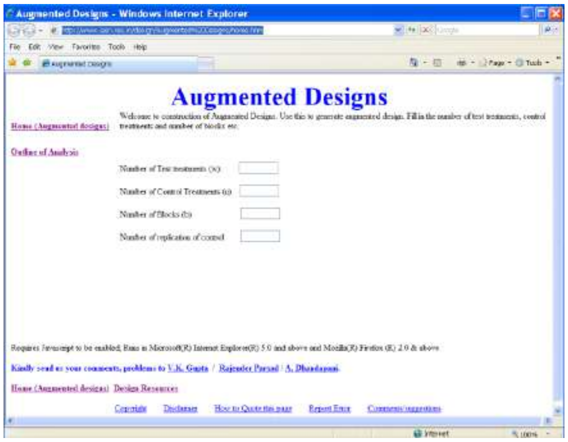
Design Resources
IASRI Home

Choose Layout: Completely Randomized Design (CRD)
Number of Treatments:
Number of Replications: In case of equal replication, enter number of replications just once. Otherwise, enter replication number of each treatment separated by comma.

Data Entry Sheet: ○ TXT ○ CSV ● No Thanks!

Kindly send us your comments, problems to Rajender Parsad / V.K. Gupta / A. Dhandapani

Copyright Disclaimer How to Quote this page Report Error Comments/suggestions



Augmented Designs - Windows Internet Explorer
File Edit View Favorites Tools Help

Augmented Designs

Home (Augmented designs)
Outline of Analysis

Welcome to construction of Augmented Designs. Use this to generate augmented design. Fill in the number of test treatments, control treatments and number of blocks etc.

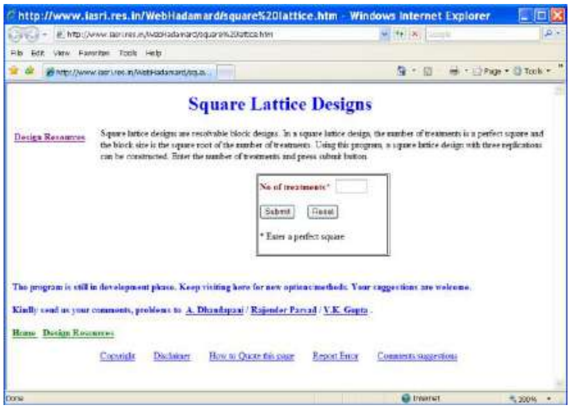
Number of Test treatments (v)
Number of Control Treatments (u)
Number of Blocks (b)
Number of replication of control

Requires Javascript to be enabled; Runs in Microsoft(R) Internet Explorer(R) 5.0 and above and Mozilla(R) Firefox (R) 2.0 & above

Kindly send us your comments, problems to V.K. Gupta / Rajender Parsad / A. Dhandapani.

Home (Augmented designs) Design Resources

Copyright Disclaimer How to Quote this page Report Error Comments/suggestions



http://www.iasri.res.in/WebHadamard/square%20lattice.htm - Windows Internet Explorer
File Edit View Favorites Tools Help

Square Lattice Designs

Design Resources

Square lattice designs are resolvable block designs. In a square lattice design, the number of treatments is a perfect square and the block size is the square root of the number of treatments. Using this program, a square lattice design with three replications can be constructed. Enter the number of treatments and press submit button.

No of treatments*

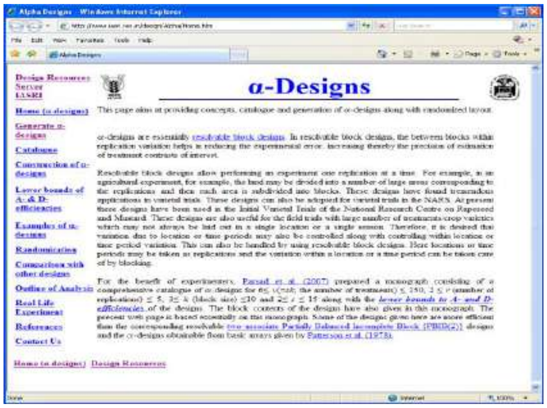
* Enter a perfect square

The program is still in development phase. Keep visiting here for new options/methods. Your suggestions are welcome.

Kindly send us your comments, problems to A. Dhandapani / Rajender Parsad / V.K. Gupta .

Home Design Resources

Copyright Disclaimer How to Quote this page Report Error Comments/suggestions



Alpha Designs - Windows Internet Explorer
File Edit View Favorites Tools Help

α-Designs

Design Resources Server IASRI
Home (α-designs)
Generate α-designs
Catalogue
Construction of α-designs
Lower bounds of A- & D-efficiencies
Examples of α-designs
Randomization
Comparison with other designs
Outline of Analysis
Real Life Experiment
References
Contact Us

This page aims at providing concepts, catalogue and generation of α-designs along with randomized layout.

α-designs are essentially resolvable block designs. In resolvable block designs, the between blocks within replication variation helps in reducing the experimental error, increasing thereby the precision of estimation of treatment contrasts of interest.

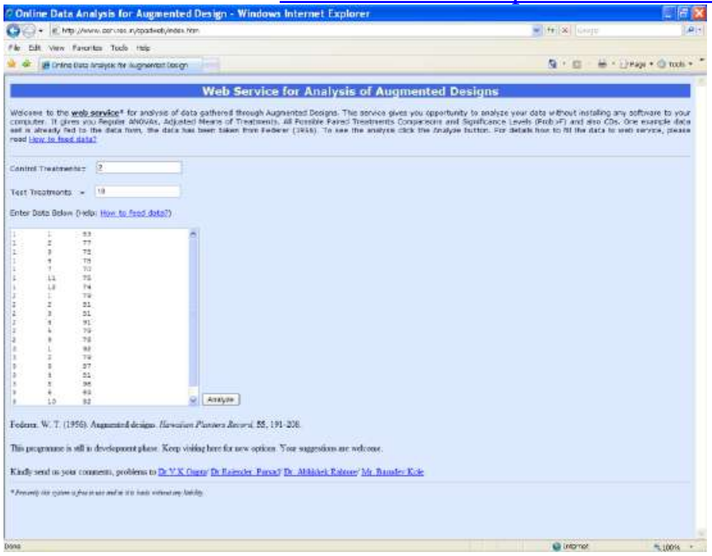
Resolvable block designs allow performing an experiment one replication at a time. For example, in an agricultural experiment, for example, the land may be divided into a number of large areas corresponding to the replications and then each area is subdivided into blocks. These designs have found tremendous applications in varietal trials. These designs can also be adopted for varietal trials in the NARS. At present these designs have been used in the Initial Varietal Trials of the National Research Centre on Rapeseed and Mustard. These designs are also useful for the field trials with large number of treatments/crop varieties which may not always be laid out in a single location or a single season. Therefore, it is desired that variation due to location or time periods may also be controlled along with controlling within location or time period variation. This can also be handled by using resolvable block designs. Here locations or time periods may be taken as replications and the variation within a location or a time period can be taken care of by blocking.

For the benefit of experimenters, Parsad et al. (2007) prepared a monograph consisting of a comprehensive catalogue of α designs for 6≤ v(=sk; the number of treatments) ≤ 150, 2 ≤ r (number of replications) ≤ 5, 3≤ k (block size) ≤10 and 2≤ s ≤ 15 along with the *lower bounds to A- and D-efficiencies* of the designs. The block contents of the designs have also given in this monograph. The present web page is based essentially on this monograph. Some of the designs given here are more efficient than the corresponding resolvable two-associate Partially Balanced Incomplete Block (PBIB(2)) designs and the α-designs obtainable from basic arrays given by Patterson et al. (1978).

Home (α-designs) Design Resources

2.3 Online Analysis of Data

This link together with Analysis of Data forms the backbone of the Design Resources Server. This particular link targets at providing online analysis of data generated to the experimenter. At present an experimenter can perform online analysis of data generated from augmented randomized block designs. This is available at www.iasri.res.in/spadweb/index.htm.



Online Data Analysis for Augmented Design - Windows Internet Explorer
File Edit View Favorites Tools Help

Web Service for Analysis of Augmented Designs

Welcome to the web service* for analysis of data gathered through Augmented Designs. The service gives you opportunity to analyze your data without installing any software to your computer. It gives you Regular ANOVAs, Adjusted Means of Treatments, All Possible Paired Treatments Comparisons and Significance Levels (Prob>F) and also CDs. One example data set is already fed to the data form, the data has been taken from Federer (1956). To see the analysis click the Analyze button. For details how to fill the data to web service, please read How to feed data?

Control Treatments: 2
Test Treatments: 18

Enter Data Below (Help: How to feed data?)

```
1    1    83
1    2    77
1    3    72
1    4    75
1    7    [illegible]
1    11   75
1    12   74
2    1    79
2    2    81
2    3    81
2    4    91
2    6    79
2    8    78
3    1    92
3    2    79
3    3    87
3    5    81
3    5    96
3    6    89
3    10   82
```

Federer, W. T. (1956). Augmented designs. *Hawaiian Planters Record*, 55, 191-208.

This programme is still in development phase. Keep visiting here for new options. Your suggestions are welcome.

Kindly send us your comments, problems to Dr V.K. Gupta/ Dr Rajender Parsad/ Dr. Abhishek Rathore/ Mr. Baudev Kole

*Presently this system is free to use and is on trial basis without any liability.

2.4 Analysis of Data

This is the most important link of the server because it targets at providing steps of analysis of data generated from designed experiments using several statistical packages like SAS, SPSS, GenStat, MINITAB, SYSTAT, SPAD, SPFE, SPAR 2.0, MS-Excel, etc. Some real life examples of experiments are given and the questions to be answered are listed. Steps for preparation of data files, the commands and macros to be used for analysis of data and the treatment contrasts to be used for answering specific questions, etc. are given, which the user can use without any difficulty. The data files and result files can also be downloaded. This is available at

www.iasri.res.in/design/Analysis of data/Analysis of Data.html.

The following analysis can be performed using this link:

- Analysis of data generated from completely randomized designs, randomized complete block design; incomplete block design; resolvable incomplete block design; Latin square design; factorial experiments both without and with confounding; factorial experiments with extra treatments; split and strip plot designs; cross over designs using SAS and SPSS; steps of analysis of augmented design using SAS, SPSS and SPAD
- Response surface design using SAS and SPSS
- SAS code for analysis of groups of experiments conducted in different environments (locations or season / year), each experiment conducted as a complete block or an incomplete block design. Using this code, one can analyze the data for each of the environments separately, test the homogeneity of error variances using Bartlett's χ^2-test, perform combined analysis of data considering both environment effects as fixed and environment effects as random (both through PROC GLM and PROC MIXED) and prepare site regression or GGE biplots
- SAS Macro for performing diagnostics (normality and homogeneity of errors) in experimental data generated through randomized complete block designs and then applying remedial measures such as Box-Cox transformation and applying the non-parametric tests if the errors remain non-normal and / or heterogeneous even after transformation
- SAS codes are also available for obtaining descriptive statistics, generating discrete frequency distribution, grouped frequency distribution, histogram, testing the normality of a given variable (overall groups or for each of the groups separately)
- correlation and regression using SAS and SPSS
- Tests of significance based on Student's *t*-distribution using SAS, SPSS and MS-EXCEL
- SAS and SPSS codes for performing principal component analysis, cluster analysis and analysis of covariance
- SAS and SPSS codes for fitting non-linear models

The screens shots for analysis of data appear like

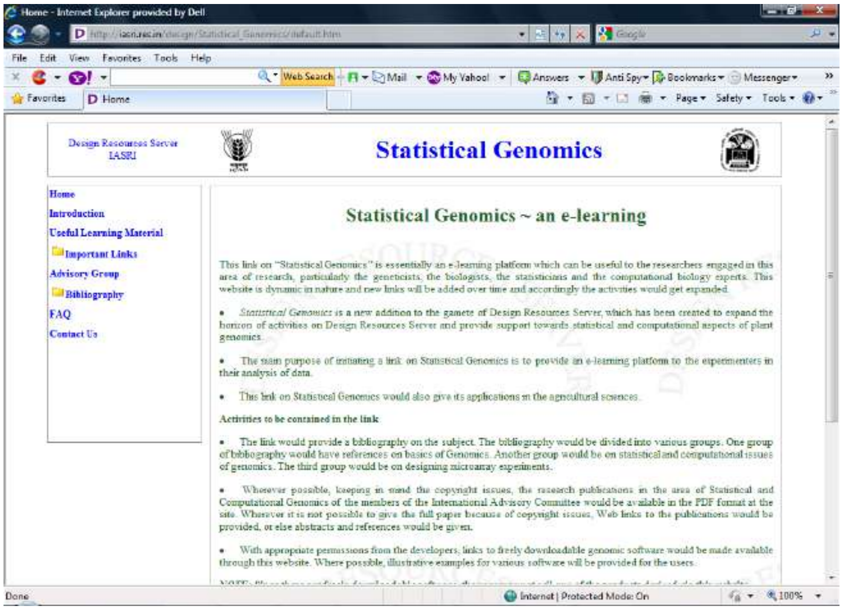
Analysis of Data from Designed Experiments

Analysis of Data from Designed Experiments

Groups of Experiments

2.6 Statistical Genomics

A link on Statistical Genomics has been initiated essentially as an e-learning platform which can be useful to the researchers particularly the geneticists, the biologists, the statisticians and the computational biology experts. It contains the information on some public domain softwares that can be downloaded free of cost. A bibliography on design and analysis of microarray experiments is also given. These are hosted at http://iasri.res.in/design/Statistical_Genomics/default.htm. A screen shot of this link is



Statistical Genomics ~ an e-learning

This link on "Statistical Genomics" is essentially an e-learning platform which can be useful to the researchers engaged in this area of research, particularly the geneticists, the biologists, the statisticians and the computational biology experts. This website is dynamic in nature and new links will be added over time and accordingly the activities would get expanded.

- *Statistical Genomics* is a new addition to the gamete of Design Resources Server, which has been created to expand the horizon of activities on Design Resources Server and provide support towards statistical and computational aspects of plant genomics.
- The main purpose of initiating a link on Statistical Genomics is to provide an e-learning platform to the experimenters in their analysis of data.
- This link on Statistical Genomics would also give its applications in the agricultural sciences.

Activities to be contained in the link

- The link would provide a bibliography on the subject. The bibliography would be divided into various groups. One group of bibliography would have references on basics of Genomics. Another group would be on statistical and computational issues of genomics. The third group would be on designing microarray experiments.
- Wherever possible, keeping in mind the copyright issues, the research publications in the area of Statistical and Computational Genomics of the members of the International Advisory Committee would be available in the PDF format at the site. Wherever it is not possible to give the full paper because of copyright issues, Web links to the publications would be provided, or else abstracts and references would be given.
- With appropriate permissions from the developers, links to freely downloadable genomic software would be made available through this website. Where possible, illustrative examples for various software will be provided for the users.

3. Useful for Research Statisticians

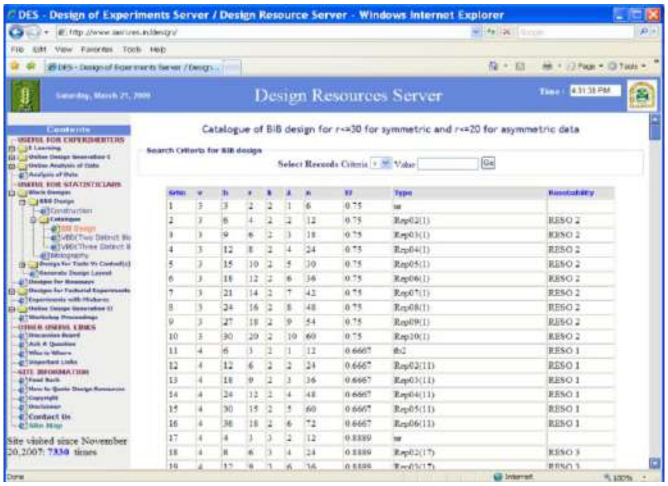
This link is useful for researchers engaged in conducting research in design of experiments and can be used for class room teaching also. The material on this link is divided into the following sub-links:

3.1 Block Designs

This link provides some theoretical considerations of balanced incomplete block (BIB) designs, binary variance balanced block (BBB) designs with 2 and 3 distinct block sizes, partially balanced incomplete block (PBIB) designs, designs for test treatments-control treatment(s) comparisons, etc. for research statisticians. The link also gives a catalogue of designs and a bibliography on the subject for use of researchers. At present the following material is available on this link:

- General method of construction of BBB designs; general methods of construction of block designs for making test treatments - control treatment(s) comparisons; bibliography
- Catalogue of BIB designs for number of replications $r \leq 30$ for symmetric BIB designs and $r \leq 20$ for asymmetric BIB designs
- Catalogue of BBB designs with 2 and 3 distinct block sizes for number of replications $r \leq 30$. The catalogue also gives the resolvability status of the designs along with the efficiency factor of the designs
- 6574 block designs for making all possible pair wise treatment comparisons for $v \leq 35$ (number of treatments), $b \leq 64$ (number of blocks), $k \leq 34$ (block size)

Some screen shots on block designs are given below:



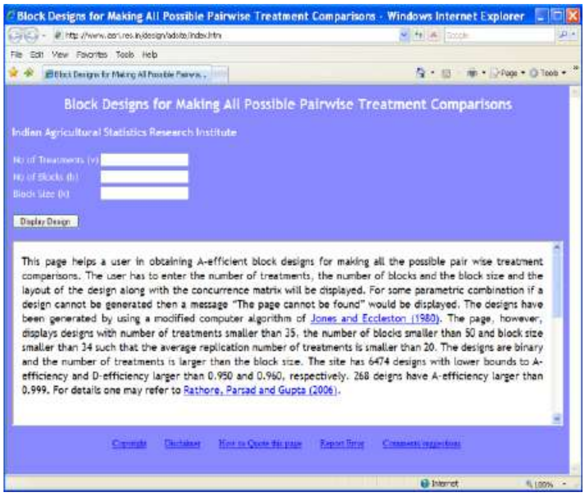
Site visited since November 20,2007: 7330 times

Catalogue of BIB design for r<=30 for symmetric and r<=20 for asymmetric data

Search Criteria for BIB design

Select Records Criteria Value Go

Srno	v	b	r	k	λ	n	E	Type	Resolvability
1	3	3	2	2	1	6	0.75	ur	
2	3	6	4	2	2	12	0.75	Rep02(1)	RESO 2
3	3	9	6	2	3	18	0.75	Rep03(1)	RESO 2
4	3	12	8	2	4	24	0.75	Rep04(1)	RESO 2
5	3	15	10	2	5	30	0.75	Rep05(1)	RESO 2
6	3	18	12	2	6	36	0.75	Rep06(1)	RESO 2
7	3	21	14	2	7	42	0.75	Rep07(1)	RESO 2
8	3	24	16	2	8	48	0.75	Rep08(1)	RESO 2
9	3	27	18	2	9	54	0.75	Rep09(1)	RESO 2
10	3	30	20	2	10	60	0.75	Rep10(1)	RESO 2
11	4	6	3	2	1	12	0.6667	fb2	RESO 1
12	4	12	6	2	2	24	0.6667	Rep02(11)	RESO 1
13	4	18	9	2	3	36	0.6667	Rep03(11)	RESO 1
14	4	24	12	2	4	48	0.6667	Rep04(11)	RESO 1
15	4	30	15	2	5	60	0.6667	Rep05(11)	RESO 1
16	4	36	18	2	6	72	0.6667	Rep06(11)	RESO 1
17	4	4	3	3	2	12	0.8889	ur	
18	4	8	6	3	4	24	0.8889	Rep02(17)	RESO 3
19	4	12	9	3	6	36	0.8889	Rep03(17)	RESO 3



Block Designs for Making All Possible Pairwise Treatment Comparisons - Windows Internet Explorer
File Edit View Favorites Tools Help
Display Design
Copyright
Disclaimer
How to Quote this page
Report Error
Comments/suggestions
Internet
100%

Block Designs for Making All Possible Pairwise Treatment Comparisons

Indian Agricultural Statistics Research Institute

No of Treatments (v)
No of Blocks (b)
Block Size (k)

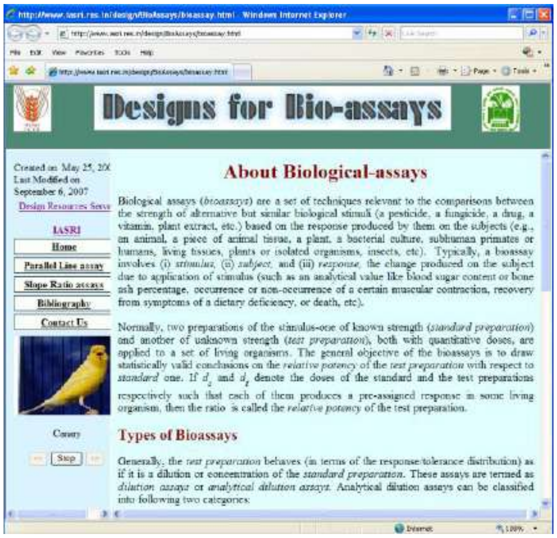
This page helps a user in obtaining A-efficient block designs for making all the possible pair wise treatment comparisons. The user has to enter the number of treatments, the number of blocks and the block size and the layout of the design along with the concurrence matrix will be displayed. For some parametric combination if a design cannot be generated then a message "The page cannot be found" would be displayed. The designs have been generated by using a modified computer algorithm of Jones and Eccleston (1980). The page, however, displays designs with number of treatments smaller than 35, the number of blocks smaller than 50 and block size smaller than 34 such that the average replication number of treatments is smaller than 20. The designs are binary and the number of treatments is larger then the block size. The site has 6474 designs with lower bounds to A-efficiency and D-efficiency larger than 0.950 and 0.960, respectively. 268 deigns have A-efficiency larger than 0.999. For details one may refer to Rathore, Parsad and Gupta (2006).

3.2 Designs for Bioassays

Designs for biological assays help in estimation of relative potency of the test preparation with respect to the standard one. The material uploaded includes contrasts of interest in parallel line assays and slope ratio assays. This link provides some theoretical considerations of designs for bioassays along with a catalogue of designs and a bibliography on the subject for use of researchers. Literature on bioassays is available at

www.iasri.res.in/design/BioAssays/bioassay.html.

Some screen shots of this link are displayed below:



Created on May 25, 20
Last Modified on September 6, 2007

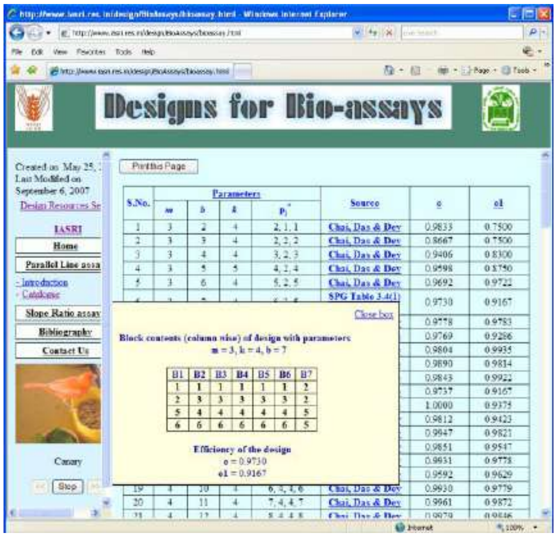
About Biological-assays

Biological assays (*bioassays*) are a set of techniques relevant to the comparisons between the strength of alternative but similar biological stimuli (a pesticide, a fungicide, a drug, a vitamin, plant extract, etc.) based on the response produced by them on the subjects (e.g., an animal, a piece of animal tissue, a plant, a bacterial culture, subhuman primates or humans, living tissues, plants or isolated organisms, insects, etc). Typically, a bioassay involves (i) *stimulus*, (ii) *subject*, and (iii) *response*, the change produced on the subject due to application of stimulus (such as an analytical value like blood sugar content or bone ash percentage, occurrence or non-occurrence of a certain muscular contraction, recovery from symptoms of a dietary deficiency, or death, etc).

Normally, two preparations of the stimulus-one of known strength (*standard preparation*) and another of unknown strength (*test preparation*), both with quantitative doses, are applied to a set of living organisms. The general objective of the bioassays is to draw statistically valid conclusions on the *relative potency* of the *test preparation* with respect to *standard* one. If d_s and d_t denote the doses of the standard and the test preparations respectively such that each of them produces a pre-assigned response in some living organism, then the ratio is called the *relative potency* of the test preparation.

Types of Bioassays

Generally, the *test preparation* behaves (in terms of the response/tolerance distribution) as if it is a dilution or concentration of the *standard preparation*. These assays are termed as *dilution assays* or *analytical dilution assays*. Analytical dilution assays can be classified into following two categories:



Created on May 25,
Last Modified on September 6, 2007

S.No.	Parameters				Source	e	e1
	m	b	k	p_i^*			
1	3	2	4	2, 1, 1	Chai, Das & Dey	0.9833	0.7500
2	3	3	4	2, 2, 2	Chai, Das & Dey	0.8667	0.7500
3	3	4	4	3, 2, 3	Chai, Das & Dey	0.9406	0.8300
4	3	5	5	4, 2, 4	Chai, Das & Dey	0.9598	0.8750
5	3	6	4	5, 2, 5	Chai, Das & Dey	0.9692	0.9722
					SPG Table 3.4(1)	0.9730	0.9167
						0.9778	0.9783
						0.9769	0.9286
						0.9804	0.9935
						0.9890	0.9814
						0.9843	0.9922
						0.9737	0.9167
						1.0000	0.9375
						0.9812	0.9423
						0.9847	0.9821
						0.9851	0.9547
						0.9931	0.9778
						0.9592	0.9629
19	4	10	4	6, 4, 4, 6	Chai, Das & Dey	0.9930	0.9779
20	4	11	4	7, 4, 4, 7	Chai, Das & Dey	0.9961	0.9872
21	4	12	4	8, 4, 4, 8	Chai, Das & Dey	0.9979	0.9846

Block contents (column wise) of design with parameters
m = 3, k = 4, b = 7

B1	B2	B3	B4	B5	B6	B7
1	1	1	1	1	1	2
2	3	3	3	3	3	2
5	4	4	4	4	4	5
6	6	6	6	6	6	5

Efficiency of the design
e = 0.9730
e1 = 0.9167

3.3 Designs for Factorial Experiments

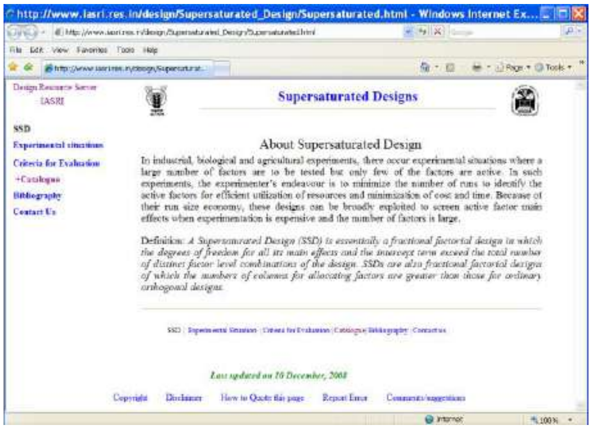
Factorial experiments are most popular among agricultural scientists. To begin with, material on block designs with factorial treatment structure and supersaturated designs is available on this link.

- **Supersaturated Designs**

Supersaturated designs are fractional factorial designs in which the degrees of freedom for all its main effects and the intercept term exceed the total number of distinct factor level combinations of the design. These designs are useful when the experimenter is interested in identifying the active factors through the experiment and experimental resources are scarce. Definition of supersaturated designs, experimental situations in which supersaturated designs are useful, efficiency criteria for evaluation of supersaturated designs, catalogue of supersaturated designs for 2-level factorial experiments and asymmetrical factorial experiments and bibliography on supersaturated designs has been uploaded on the Server. The complete details of the runs can be obtained by clicking on the required design in the catalogue.

www.iasri.res.in/design/Supersaturated_Design/Supersaturated.html.

Some screen shots of supersaturated designs are


http://www.iasri.res.in/design/Supersaturated_Design/Supersaturated.html - Windows Internet Ex...
File Edit View Favorites Tools Help
Page Tools
Internet
100%

Design Resource Server
IASRI

SSD
Experimental situations
Criteria for Evaluation
+Catalogue
Bibliography
Contact Us

Supersaturated Designs

About Supersaturated Design

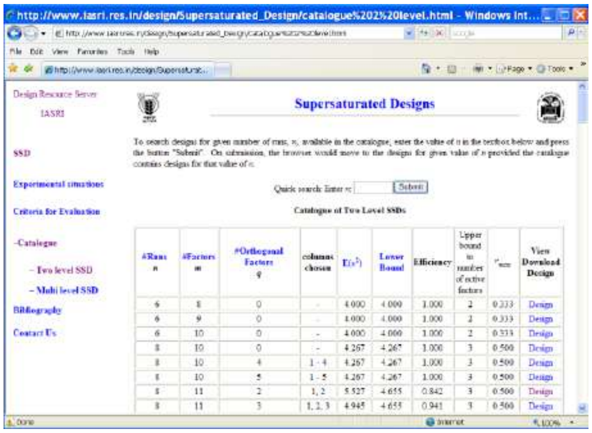
In industrial, biological and agricultural experiments, there occur experimental situations where a large number of factors are to be tested but only few of the factors are active. In such experiments, the experimenter's endeavour is to minimize the number of runs to identify the active factors for efficient utilization of resources and minimization of cost and time. Because of their run size economy, these designs can be broadly exploited to screen active factor main effects when experimentation is expensive and the number of factors is large.

Definition: *A Supersaturated Design (SSD) is essentially a fractional factorial design in which the degrees of freedom for all its main effects and the intercept term exceed the total number of distinct factor level combinations of the design. SSDs are also fractional factorial designs of which the numbers of columns for allocating factors are greater than those for ordinary orthogonal designs.*

SSD | Experimental Situation | Criteria for Evaluation | Catalogue | Bibliography | Contact us

Last updated on 10 December, 2008

Copyright Disclaimer How to Quote this page Report Error Comments/suggestions


http://www.iasri.res.in/design/Supersaturated_Design/catalogue%202%20level.html - Windows Int...
File Edit View Favorites Tools Help
Page Tools
Done
Internet
100%

Design Resource Server
IASRI

SSD
Experimental situations
Criteria for Evaluation
-Catalogue
– Two level SSD
– Multi level SSD
Bibliography
Contact Us

Supersaturated Designs

To search designs for given number of runs, *n*, available in the catalogue, enter the value of *n* in the textbox below and press the button "Submit". On submission, the browser would move to the designs for given value of *n* provided the catalogue contains designs for that value of *n*.

Quick search: Enter *n*: Submit

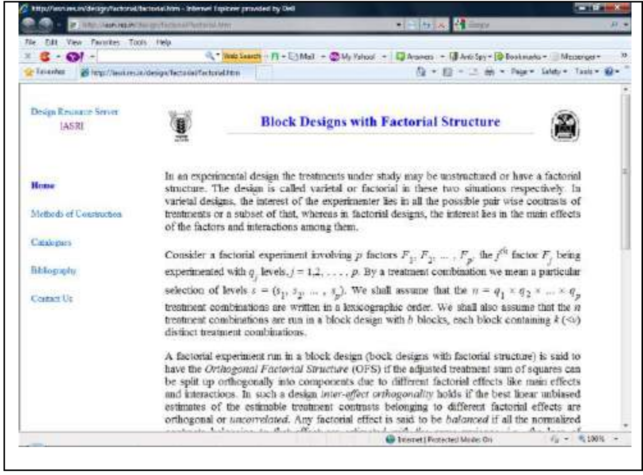
Catalogue of Two Level SSDs

#Runs *n*	#Factors *m*	#Orthogonal Factors *q*	columns chosen	$E(s^2)$	Lower Bound	Efficiency	Upper bound to number of active factors	r_{max}	View Download Design
6	8	0	-	4.000	4.000	1.000	2	0.333	Design
6	9	0	-	4.000	4.000	1.000	2	0.333	Design
6	10	0	-	4.000	4.000	1.000	2	0.333	Design
8	10	0	-	4.267	4.267	1.000	3	0.500	Design
8	10	4	1 - 4	4.267	4.267	1.000	3	0.500	Design
8	10	5	1 - 5	4.267	4.267	1.000	3	0.500	Design
8	11	2	1, 2	5.527	4.655	0.842	3	0.500	Design
8	11	3	1, 2, 3	4.945	4.655	0.941	3	0.500	Design

➢ Block Designs with Factorial Treatment Structure

Block designs with factorial treatment structure have useful applications in designs for crop sequence experiments. Th link on block designs with factorial Treatment Structure provides a bibliography with 232 references on the subject. Catalogues of block designs with factorial treatment structure in 3-replications for number of levels for any factor at most 12 permitting estimation of main effects with full efficiency and controlling efficiency for interaction effects are also given at this link. URL for this link is www.iasri.res.in/design/factorial/factorial.htm.

Some screen shots for block designs with factorial treatment structure are


http://iasri.res.in/design/factorial/factorial.htm - Internet Explorer provided by Dell
File Edit View Favorites Tools Help
Web Search
Mail
My Yahoo!
Answers
Anti Spy
Bookmarks
Messenger
Favorites
http://iasri.res.in/design/factorial/factorial.htm
Page Safety Tools
Internet | Protected Mode: On
100%

Design Resource Server
IASRI

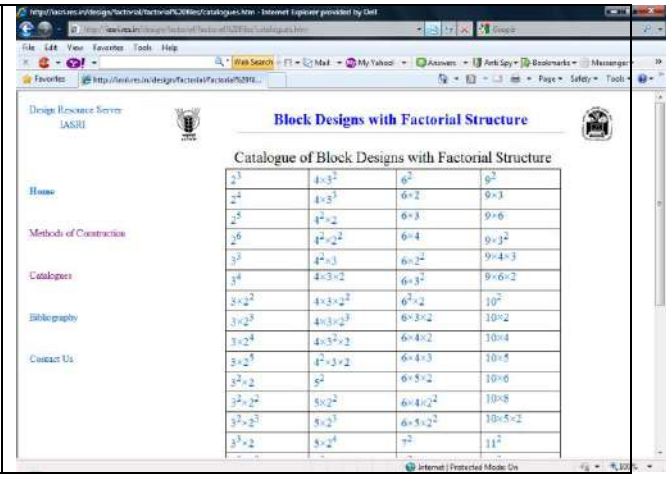
Home
Methods of Construction
Catalogues
Bibliography
Contact Us

Block Designs with Factorial Structure

In an experimental design the treatments under study may be unstructured or have a factorial structure. The design is called varietal or factorial in these two situations respectively. In varietal designs, the interest of the experimenter lies in all the possible pair wise contrasts of treatments or a subset of that, whereas in factorial designs, the interest lies in the main effects of the factors and interactions among them.

Consider a factorial experiment involving p factors $F_1, F_2, \ldots, F_p$, the j^{th} factor F_j being experimented with q_j levels, $j = 1, 2, \ldots, p$. By a treatment combination we mean a particular selection of levels $s = (s_1, s_2, \ldots, s_p)$. We shall assume that the $n = q_1 \times q_2 \times \ldots \times q_p$ treatment combinations are written in a lexicographic order. We shall also assume that the n treatment combinations are run in a block design with b blocks, each block containing k ($<v$) distinct treatment combinations.

A factorial experiment run in a block design (bock designs with factorial structure) is said to have the *Orthogonal Factorial Structure* (OFS) if the adjusted treatment sum of squares can be split up orthogonally into components due to different factorial effects like main effects and interactions. In such a design *inter-effect orthogonality* holds if the best linear unbiased estimates of the estimable treatment contrasts belonging to different factorial effects are orthogonal or *uncorrelated*. Any factorial effect is said to be *balanced* if all the normalized


http://iasri.res.in/design/factorial/factorial%20files/catalogues.htm - Internet Explorer provided by Dell
File Edit View Favorites Tools Help
Web Search
Mail
My Yahoo!
Answers
Anti Spy
Bookmarks
Messenger
Favorites
http://iasri.res.in/design/Factorial/Factorial%20fil...
Page Safety Tools
Internet | Protected Mode: On

Design Resource Server
IASRI

Home
Methods of Construction
Catalogues
Bibliography
Contact Us

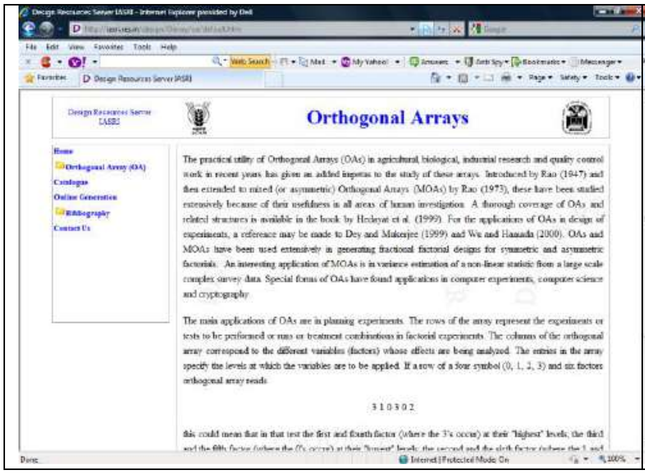
Block Designs with Factorial Structure

Catalogue of Block Designs with Factorial Structure

2^3	4×3^2	6^2	9^2
2^4	4×3^3	6×2	9×3
2^5	$4^2\times2$	6×3	9×6
2^6	$4^2\times2^2$	6×4	9×3^2
3^3	$4^2\times3$	6×2^2	$9\times4\times3$
3^4	$4\times3\times2$	6×3^2	$9\times6\times2$
3×2^2	$4\times3\times2^2$	$6^2\times2$	10^2
3×2^3	$4\times3\times2^3$	$6\times3\times2$	10×2
3×2^4	$4\times3^2\times2$	$6\times4\times2$	10×4
3×2^5	$4^2\times3\times2$	$6\times4\times3$	10×5
$3^2\times2$	5^2	$6\times5\times2$	10×6
$3^2\times2^2$	5×2^2	$6\times4\times2^2$	10×8
$3^2\times2^3$	5×2^3	$6\times5\times2^2$	$10\times5\times2$
$3^3\times2$	5×2^4	7^2	11^2

➢ Mixed Orthogonal arrays

Definitions of Orthogonal arrays(OAs), mixed OA, Resolvable OA, α-resolvable OA, resolvable MOA, construction of OAs, blocking in OAs, generation of orthogonal arrays of strength two, resolvable orthogonal arrays of strength two and the orthogonal blocking of the resolvable orthogonal array for $4 \leq n$(# Runs) ≤ 144, and bibliography on OAs.


Design Resources Server IASRI - Internet Explorer provided by Dell
File Edit View Favorites Tools Help
Web Search
Mail
My Yahoo!
Answers
Anti Spy
Bookmarks
Messenger
Favorites
Design Resources Server IASRI
Page
Safety
Tools
Done
Internet | Protected Mode: On
100%

Design Resources Server IASRI

Orthogonal Arrays

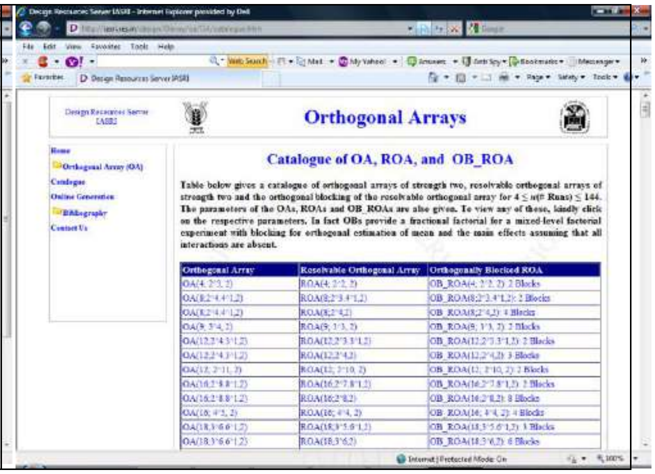
Home
Orthogonal Array (OA)
Catalogue
Online Generation
Bibliography
Contact Us

The practical utility of Orthogonal Arrays (OAs) in agricultural, biological, industrial research and quality control work in recent years has given an added impetus to the study of these arrays. Introduced by Rao (1947) and then extended to mixed (or asymmetric) Orthogonal Arrays (MOAs) by Rao (1973), these have been studied extensively because of their usefulness in all areas of human investigation. A thorough coverage of OAs and related structures is available in the book by Hedayat et al. (1999). For the applications of OAs in design of experiments, a reference may be made to Dey and Mukerjee (1999) and Wu and Hamada (2000). OAs and MOAs have been used extensively in generating fractional factorial designs for symmetric and asymmetric factorials. An interesting application of MOAs is in variance estimation of a non-linear statistic from a large scale complex survey data. Special forms of OAs have found applications in computer experiments, computer science and cryptography.

The main applications of OAs are in planning experiments. The rows of the array represent the experiments or tests to be performed or runs or treatment combinations in factorial experiments. The columns of the orthogonal array correspond to the different variables (factors) whose effects are being analyzed. The entries in the array specify the levels at which the variables are to be applied. If a row of a four symbol (0, 1, 2, 3) and six factors orthogonal array reads

3 1 0 3 0 2

this could mean that in that test the first and fourth factor (where the 3's occur) at their "highest" levels, the third and the fifth factor (where the 0's occur) at their "lowest" levels, the second and the sixth factor (where the 1 and


Design Resources Server IASRI - Internet Explorer provided by Dell
File Edit View Favorites Tools Help
Web Search
Mail
My Yahoo!
Answers
Anti Spy
Bookmarks
Messenger
Favorites
Design Resources Server IASRI
Page
Safety
Tools
Internet | Protected Mode: On
100%

Design Resources Server IASRI

Orthogonal Arrays

Home
Orthogonal Array (OA)
Catalogue
Online Generation
Bibliography
Contact Us

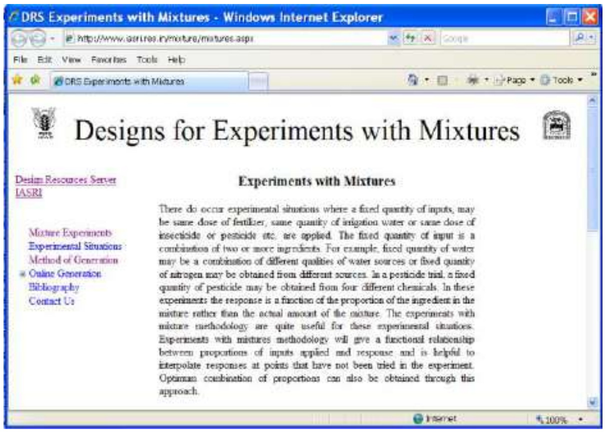
Catalogue of OA, ROA, and OB_ROA

Table below gives a catalogue of orthogonal arrays of strength two, resolvable orthogonal arrays of strength two and the orthogonal blocking of the resolvable orthogonal array for 4 ≤ n(# Runs) ≤ 144. The parameters of the OAs, ROAs and OB_ROAs are also given. To view any of these, kindly click on the respective parameters. In fact OBs provide a fractional factorial for a mixed-level factorial experiment with blocking for orthogonal estimation of mean and the main effects assuming that all interactions are absent.

Orthogonal Array	Resolvable Orthogonal Array	Orthogonally Blocked ROA
OA(4; 2^3, 2)	ROA(4; 2^2, 2)	OB_ROA(4; 2^2, 2): 2 Blocks
OA(8;2^4.4^1,2)	ROA(8;2^3.4^1,2)	OB_ROA(8;2^3.4^1,2): 2 Blocks
OA(8;2^4.4^1,2)	ROA(8;2^4,2)	OB_ROA(8;2^4,2): 4 Blocks
OA(9; 3^4, 2)	ROA(9; 3^3, 2)	OB_ROA(9; 3^3, 2): 2 Blocks
OA(12;2^4.3^1,2)	ROA(12;2^3.3^1,2)	OB_ROA(12;2^3.3^1,2): 2 Blocks
OA(12;2^4.3^1,2)	ROA(12;2^4,2)	OB_ROA(12;2^4,2): 3 Blocks
OA(12; 2^11, 2)	ROA(12; 2^10, 2)	OB_ROA(12; 2^10, 2): 2 Blocks
OA(16;2^8.8^1,2)	ROA(16;2^7.8^1,2)	OB_ROA(16;2^7.8^1,2): 2 Blocks
OA(16;2^8.8^1,2)	ROA(16;2^8,2)	OB_ROA(16;2^8,2): 8 Blocks
OA(16; 4^5, 2)	ROA(16; 4^4, 2)	OB_ROA(16; 4^4, 2): 4 Blocks
OA(18;3^6.6^1,2)	ROA(18;3^5.6^1,2)	OB_ROA(18;3^5.6^1,2): 3 Blocks
OA(18;3^6.6^1,2)	ROA(18;3^6,2)	OB_ROA(18;3^6,2): 6 Blocks

3.4 Experiments with Mixtures

Experiments with mixtures are quite useful for the experiments where a fixed quantity of inputs (may be same dose of fertilizer, same quantity of irrigation water or same dose of insecticide or pesticide etc.) are applied as a combination of two or more ingredients. In these experiments the response is a function of the proportion of the ingredients in the mixture rather than the actual amount of the mixture. A bibliography of experiments with mixtures and online generation of simplex centroid designs are available on this page http://www.iasri.res.in/mixture/mixtures.aspx. Some screen shots of experiments with mixtures are:


DRS Experiments with Mixtures - Windows Internet Explorer
http://www.iasri.res.in/mixture/mixtures.aspx
File Edit View Favorites Tools Help
DRS Experiments with Mixtures
Page
Tools
Internet
100%

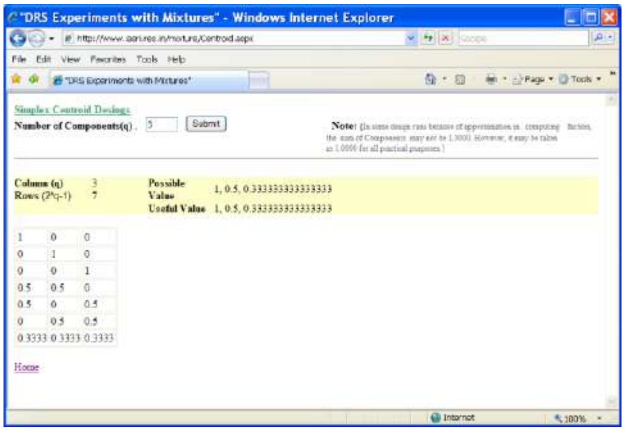
Designs for Experiments with Mixtures

Design Resources Server
IASRI

Mixture Experiments
Experimental Situations
Method of Generation
Online Generation
Bibliography
Contact Us

Experiments with Mixtures

There do occur experimental situations where a fixed quantity of inputs, may be same dose of fertilizer, same quantity of irrigation water or same dose of insecticide or pesticide etc. are applied. The fixed quantity of input is a combination of two or more ingredients. For example, fixed quantity of water may be a combination of different qualities of water sources or fixed quantity of nitrogen may be obtained from different sources. In a pesticide trial, a fixed quantity of pesticide may be obtained from four different chemicals. In these experiments the response is a function of the proportion of the ingredient in the mixture rather than the actual amount of the mixture. The experiments with mixture methodology are quite useful for these experimental situations. Experiments with mixtures methodology will give a functional relationship between proportions of inputs applied and response and is helpful to interpolate responses at points that have not been tried in the experiment. Optimum combination of proportions can also be obtained through this approach.


"DRS Experiments with Mixtures" - Windows Internet Explorer
http://www.iasri.res.in/mixture/Centroid.aspx
File Edit View Favorites Tools Help
"DRS Experiments with Mixtures"
Page
Tools
Submit
Internet
100%

Simplex Centroid Designs

Number of Components(q) : 3

Note: (In some design runs because of approximation in computing fraction, the sum of Components may not be 1.0000. However, it may be taken as 1.0000 for all practical purposes.)

Column (q) 3
Rows (2^q-1) 7

Possible Value 1, 0.5, 0.333333333333333
Useful Value 1, 0.5, 0.333333333333333

1	0	0
0	1	0
0	0	1
0.5	0.5	0
0.5	0	0.5
0	0.5	0.5
0.3333	0.3333	0.3333

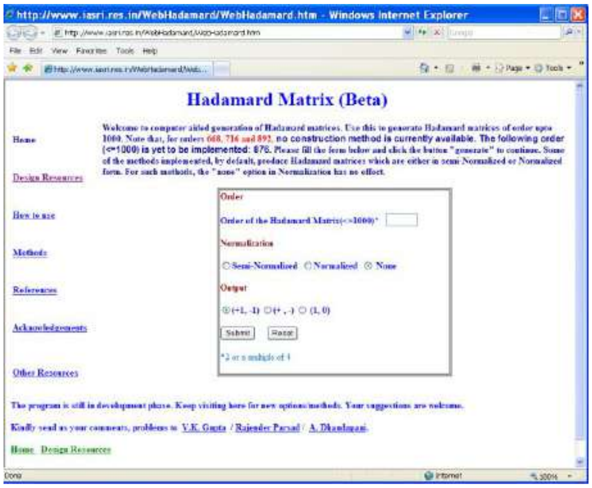
Home

3.5 Online Design Generation- II

This link is helpful in generation of the following:

Hadamard Matrix

Hadamard matrices have a tremendous potential for applications in many fields particularly in fractional factorial plans, supersaturated designs, variance estimation from large scale complex survey data, generation of incomplete block designs, coding theory, etc. One can generate Hadamard matrices for all permissible orders up to 1000 except 668, 716, 876 and 892 using the URL www.iasri.res.in/WebHadamard/WebHadamard.htm. Methods implemented produce Hadamard matrices in semi-normalized or normalized form. “None” option is also available. Hadamard matrix can be generated in (0,1); (+1,-1); or (+,-) form. The method of generation of Hadamard matrix is also given. The screen shots for generation of Hadamard matrices are



http://www.iasri.res.in/WebHadamard/WebHadamard.htm - Windows Internet Explorer

Hadamard Matrix (Beta)

Welcome to computer aided generation of Hadamard matrices. Use this to generate Hadamard matrices of order upto 1000. Note that, for orders 668, 716 and 892, no construction method is currently available. The following order (<=1000) is yet to be implemented: 876. Please fill the form below and click the button "generate" to continue. Some of the methods implemented, by default, produce Hadamard matrices which are either in semi Normalized or Normalized form. For such methods, the "none" option in Normalization has no effect.

Order

Order of the Hadamard Matrix(<=1000)* ▭

Normalization

○ Semi-Normalized ○ Normalized ● None

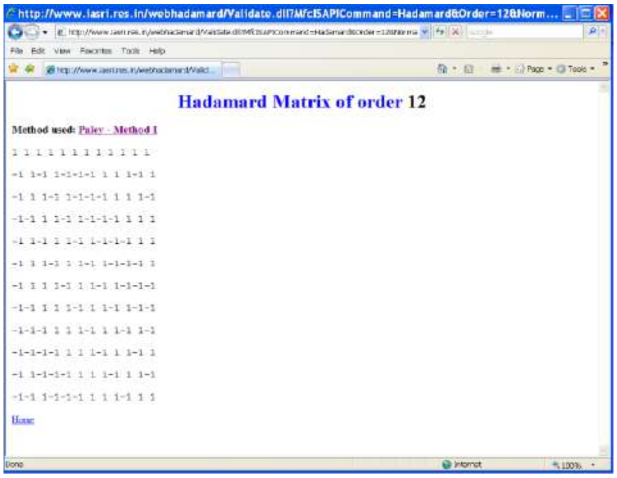
Output

● (+1, -1) ○ (+ , -) ○ (1, 0)

*2 or a multiple of 4

The program is still in development phase. Keep visiting here for new options/methods. Your suggestions are welcome.

Kindly send us your comments, problems to V.K. Gupta / Rajender Parsad / A. Dhandapani.



http://www.iasri.res.in/webhadamard/Validate.dll?MfcISAPICommand=Hadamard&Order=12&Norm...

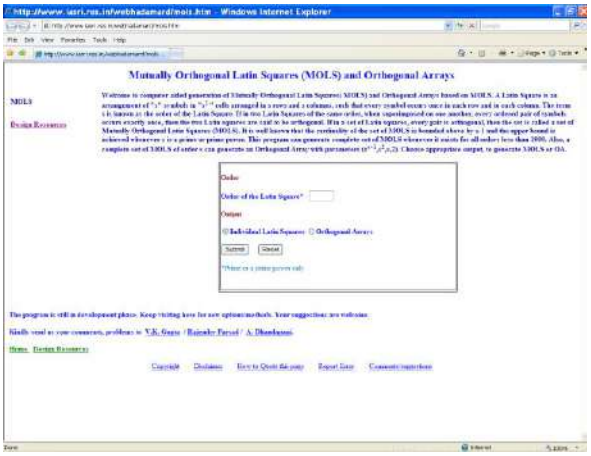
Hadamard Matrix of order 12

Method used: Paley - Method I

```
1 1 1 1 1 1 1 1 1 1 1 1
-1 1-1 1-1-1-1 1 1 1-1 1
-1 1 1-1 1-1-1-1 1 1 1-1
-1-1 1 1-1 1-1-1-1 1 1 1
-1 1-1 1 1-1 1-1-1-1 1 1
-1 1 1-1 1 1-1 1-1-1-1 1
-1 1 1 1-1 1 1-1 1-1-1-1
-1-1 1 1 1-1 1 1-1 1-1-1
-1-1-1 1 1 1-1 1 1-1 1-1
-1-1-1-1 1 1 1-1 1 1-1 1
-1 1-1-1-1 1 1 1-1 1 1-1
-1-1 1-1-1-1 1 1 1-1 1 1
```

Mutually Orthogonal Latin Squares and Orthogonal arrays

Using this link one can generate complete set of mutually orthogonal Latin squares of order s, s being a prime or prime power less than 1000. One can also generate an orthogonal array with parameters (s^{s+1}, s^2, s, 2) by choosing the output option as orthogonal arrays. The URL of this link is www.iasri.res.in/WebHadamard/mols.htm. Some screen shots of mutually orthogonal Latin squares and orthogonal arrays are



http://www.iasri.res.in/webhadamard/mols.htm - Windows Internet Explorer

Mutually Orthogonal Latin Squares (MOLS) and Orthogonal Arrays

Welcome to computer aided generation of Mutually Orthogonal Latin Squares (MOLS) and Orthogonal Arrays based on MOLS. A Latin Square is an arrangement of "s" symbols in "s^2" cells arranged in s rows and s columns, such that every symbol occurs once in each row and in each column. The term s is known as the order of the Latin Square. If in two Latin Squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, then the two Latin squares are said to be orthogonal. If in a set of Latin squares, every pair is orthogonal, then the set is called a set of Mutually Orthogonal Latin Squares (MOLS). It is well known that the cardinality of the set of MOLS is bounded above by s-1 and the upper bound is achieved whenever s is a prime or prime power. This program can generate complete set of MOLS whenever it exists for all orders less than 1000. Also, a complete set of MOLS of order s can generate an Orthogonal Array with parameters (s^{s+1},s^2,s,2). Choose appropriate output, to generate MOLS or OA.

Order

Order of the Latin Square* ▭

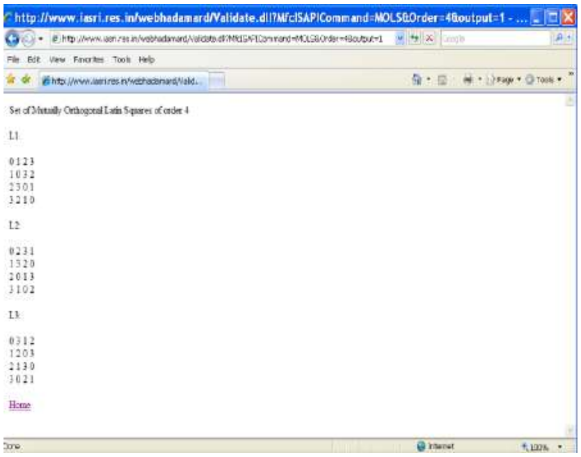
Output

● Individual Latin Squares ○ Orthogonal Arrays

*Prime or a prime power only

The program is still in development phase. Keep visiting here for new options/methods. Your suggestions are welcome.

Kindly send us your comments, problems to V.K. Gupta / Rajender Parsad / A. Dhandapani.



http://www.iasri.res.in/webhadamard/Validate.dll?MfcISAPICommand=MOLS&Order=4&output=1 - ...

Set of Mutually Orthogonal Latin Squares of order 4

```
L1

0 1 2 3
1 0 3 2
2 3 0 1
3 2 1 0

L2

0 2 3 1
1 3 2 0
2 0 1 3
3 1 0 2

L3

0 3 1 2
1 2 0 3
2 1 3 0
3 0 2 1
```

3.6 Workshop Proceedings

Proceedings of 3 dissemination workshops are available for the stakeholders

1. Design and Analysis of On-Station and On-Farm Agricultural Experiments
2. Design and Analysis of Bioassays
3. Outliers in Designed Experiments

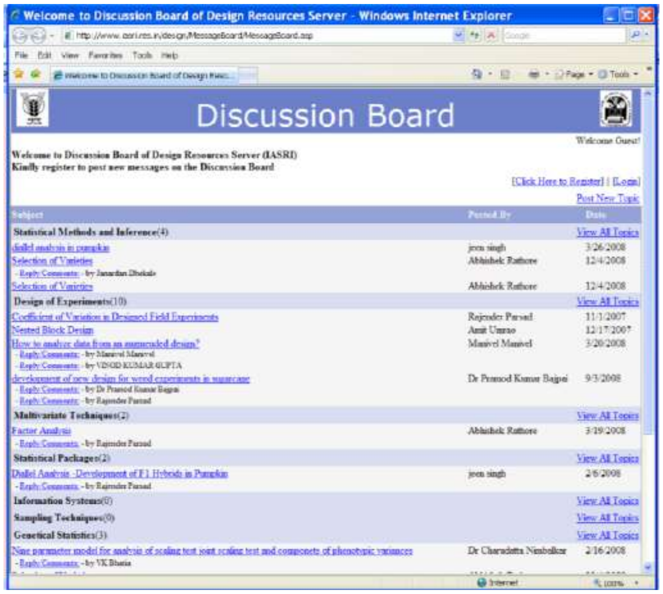
4. Other Useful Links

The purpose of this component is to develop a network of scientists in general and a network of statisticians in particular around the globe so that interesting and useful information can be shared among the peers. It also attempts to provide a sort of advisory to the scientists. Some other useful and important links available on world wide web are also provided.

4.1 Discussion Board

The purpose of discussion board is to create a network of scientists and also to provide a platform for sharing any useful piece of research or idea with scientists over the globe. The user can use this board for learning and disseminating information after registering on the discussion board.

The information can be viewed by anybody over the globe. In case there are some queries or some researchable issues, then other peers can also respond to these queries. This helps in creating a network of scientists. Number of registered participants so far is 78 (23: Agricultural Research Statisticians; 37: Experimenters; One Vice-Chancellor and 17 ISS Officers). (www.iasri.res.in/design/MessageBoard/MessageBoard.asp).


Welcome to Discussion Board of Design Resources Server - Windows Internet Explorer
http://www.iasri.res.in/design/MessageBoard/MessageBoard.asp
File Edit View Favorites Tools Help
Welcome to Discussion Board of Design Reso...
Page Tools
Internet
100%

Discussion Board

Welcome Guest!

Welcome to Discussion Board of Design Resources Server (IASRI)
Kindly register to post new messages on the Discussion Board

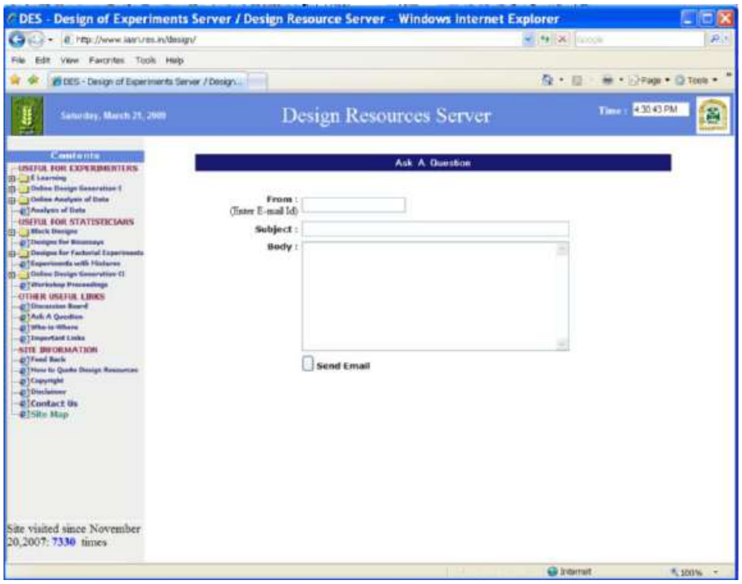
[Click Here to Register] | [Login]

Post New Topic

Subject	Posted By	Date
Statistical Methods and Inference(4)		View All Topics
diallel analysis in pumpkin	jeen singh	3/26/2008
Selection of Varieties	Abhishek Rathore	12/4/2008
- Reply Comments: - by Janardan Dhakale		
Selection of Varieties	Abhishek Rathore	12/4/2008
Design of Experiments(10)		View All Topics
Coefficient of Variation in Designed Field Experiments	Rajender Parsad	11/1/2007
Nested Block Design	Amit Umrao	12/17/2007
How to analyze data from an augmented design?	Manivel Manivel	3/20/2008
- Reply Comments: - by Manivel Manivel		
- Reply Comments: - by VINOD KUMAR GUPTA		
development of new design for weed experiments in sugarcane	Dr Pramod Kumar Bajpai	9/3/2008
- Reply Comments: - by Dr Pramod Kumar Bajpai		
- Reply Comments: - by Rajender Parsad		
Multivariate Techniques(2)		View All Topics
Factor Analysis	Abhishek Rathore	3/19/2008
- Reply Comments: - by Rajender Parsad		
Statistical Packages(2)		View All Topics
Diallel Analysis -Development of F1 Hybrids in Pumpkin	jeen singh	2/6/2008
- Reply Comments: - by Rajender Parsad		
Information Systems(0)		View All Topics
Sampling Techniques(0)		View All Topics
Genetical Statistics(3)		View All Topics
Nine parameter model for analysis of scaling test joint scaling test and componets of phenotypic variances	Dr Charudatta Nimbalkar	2/16/2008
- Reply Comments: - by VK Bhatia		

4.2 Ask a Question

The ultimate objective of this server is to provide e-learning and e-advisory services. At present this is being achieved through the link "Ask a Question". Once a user submits a question, a mail is automatically generated for Dr. Rajender Parsad, Dr. V.K. Gupta and Mrs. Alka Arora, who answer the question on receiving the mail.


DES - Design of Experiments Server / Design Resource Server - Windows Internet Explorer
http://www.iasri.res.in/design/
File Edit View Favorites Tools Help
DES - Design of Experiments Server / Design...
Page Tools
Contents
USEFUL FOR EXPERIMENTERS
E Learning
Online Design Generation I
Online Analysis of Data
Analysis of Data
USEFUL FOR STATISTICIANS
Block Designs
Designs for Bioassays
Designs for Factorial Experiments
Experiments with Mixtures
Online Design Generation II
Workshop Proceedings
OTHER USEFUL LINKS
Discussion Board
Ask A Question
Who-is-Where
Important Links
SITE INFORMATION
Feed Back
How to Quote Design Resources
Copyright
Disclaimer
Contact Us
Site Map
Internet
100%

Saturday, March 21, 2009

Design Resources Server

Time : 4:30:43 PM

Ask A Question

From :
(Enter E-mail Id)

Subject :

Body :

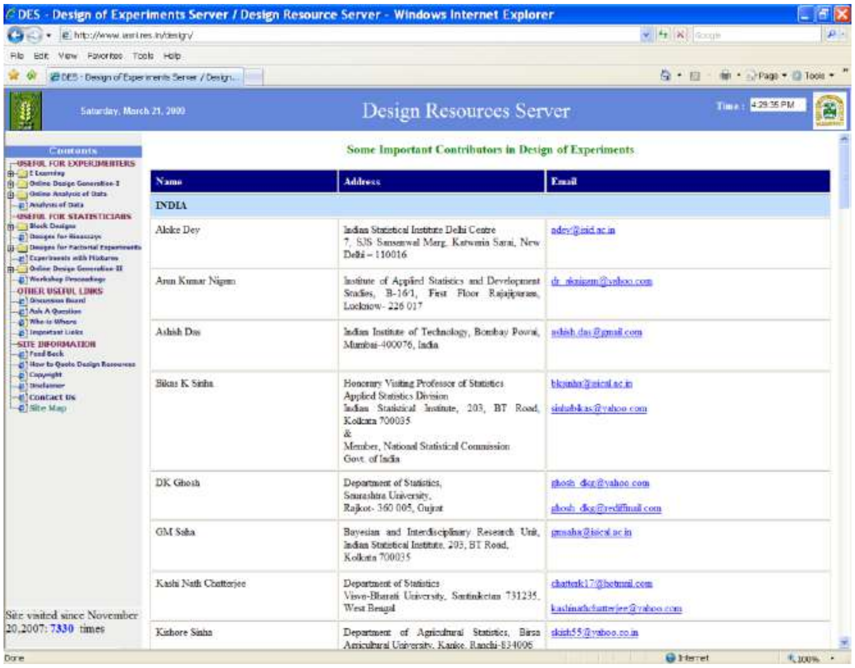
□ Send Email

Site visited since November 20,2007: 7330 times

4.3 Who-is-where

Addresses of important contributors in Design of Experiments including their E-mail addresses have been linked to Design Resources Server. The list includes experts from USA, Canada, Australia, UK, China, Japan, Mexico, New Zealand, Oman, Syria, Taiwan, Vietnam and India.

This information is useful for all the researchers in Design of Experiments in establishing linkages with their counterparts over the globe.



Some Important Contributors in Design of Experiments

Name	Address	Email
INDIA		
Aloke Dey	Indian Statistical Institute Delhi Centre 7, SJS Sansanwal Marg, Katwaria Sarai, New Delhi – 110016	adey@isid.ac.in
Arun Kumar Nigam	Institute of Applied Statistics and Development Studies, B-16/1, First Floor Rajajipuram, Lucknow- 226 017	dr_aknigam@yahoo.com
Ashish Das	Indian Institute of Technology, Bombay Powai, Mumbai-400076, India	ashish.das@gmail.com
Bikas K Sinha	Honorary Visiting Professor of Statistics Applied Statistics Division Indian Statistical Institute, 203, BT Road, Kolkata 700035 & Member, National Statistical Commission Govt. of India	bksinha@isical.ac.in sinhabikas@yahoo.com
DK Ghosh	Department of Statistics, Saurashtra University, Rajkot- 360 005, Gujrat	ghosh_dkg@yahoo.com ghosh_dkg@rediffmail.com
GM Saha	Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, 203, BT Road, Kolkata 700035	gmsaha@isical.ac.in
Kashi Nath Chatterjee	Department of Statistics Visva-Bharati University, Santiniketan 731235, West Bengal	chatterk17@hotmail.com kashinathchatterjee@yahoo.com
Kishore Sinha	Department of Agricultural Statistics, Birsa Agricultural University, Kanke, Ranchi-834006	skish55@yahoo.co.in

Site visited since November 20,2007: 7330 times

4.4 Important Links

This gives links to other important sites that provide useful material on statistical learning in general and Design of Experiments in particular. Some links are as given below:

S No.	Important Links
1.	Design Resources: www.designtheory.org
2.	Statistics Glossary http://www.cas.lancs.ac.uk/glossary_v1.1/main.html
3.	Free Encyclopedia on Design of Experiments: http://en.wikipedia.org/wiki/Design_of_experiments
4.	Important Contributors to Statistics: http://en.wikipedia.org/wiki/Statistics#Important_contributors_to_statistics
5.	Electronic Statistics Text Book: http://www.statsoft.com/textbook/stathome.html
6.	On-line construction of Designs: http://biometrics.hri.ac.uk/experimentaldesigns/website/hri.htm
7.	GENDEX: http://www.designcomputing.net/gendex/
8.	Hadamard Matrices *1. http://www.research.att.com/~njas/hadamard* 2. http://www.uow.edu.au/~jennie/WILLIAMSON/williamson.html
9.	Biplots :http://www.ggebiplot.com
10.	Free Statistical Softwares: http://freestatistics.altervista.org/en/stat.php
11.	Learning Statistics: http://freestatistics.altervista.org/en/learning.php
12.	Statistical Calculators: http://www.graphpad.com/quickcalcs/index.cfm
13.	SAS Online Doc 9.1.3: http://support.sas.com/onlinedoc/913/docMainpage.jsp
14.	University of South California: Courses in Statistics: http://www.stat.sc.edu/curricula/courses/

15.	Course on Introduction to Experimental Design: http://www.stat.sc.edu/~grego/courses/stat506
16.	Course on Experimental design: http://www.stat.sc.edu/~grego/courses/stat706

5. Site Information

This link provides information about the site on the following aspects (i) Feedback from stakeholders, (ii) How to Quote Design Resources Server, (iii) Copyright, (iv) Disclaimer, (v) Contact us, and (vi) Sitemap.

5.1 Feedback/ Comments

The feedback / comments received from the users visiting the site have been put on the server so that every user can benefit from the experience of other users. More importantly, the feedback helps in improving the contents of the site and their presentation too. We have received feedback from 19 researchers (6: Design Experts from India; 7: Experts from abroad; 4: Experimenters and 2: Agricultural Research Statisticians). The first feedback was received from Dr K Rameash, Entomologist working at ICAR Research Complex for NEH Region, Sikkim Centre, Tadong, Gangtok.

5.2 How to quote Design Resources Server

To Quote Design Resources Server, use:

Design Resources Server. *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India*. www.iasri.res.in/design (accessed last on <date>).

If referring to a particular page, then the site may be quoted as

Authors' name in 'Contact Us' list on that page. Title of page: Design Resources Server. *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India*. www.iasri.res.in/design (accessed last on <date>).

For example, page on alpha designs may be cited as
Parsad, R., Gupta, V.K. and Dhandapani, A. Alpha Designs: Design Resources Server. *Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India*. www.iasri.res.in/design (accessed last on 21.03.2009).

5.3 Copyright

This website and its contents are copyright of "IASRI (ICAR)" - © "ICAR" 2008. All rights reserved. Any redistribution or reproduction of part or all of the contents in any form, other than the following, is prohibited:

- print or download to a local hard disk extracts for personal and non-commercial use only.
- transmit it or store it in any other website or other form of electronic retrieval system.
- except with express written permission of the authors, distribution or commercial exploitation of the contents.

5.4 Disclaimer

The information contained in this website is for general information purposes only. The information is provided by “IASRI” and whilst “IASRI” endeavours to keep the information up-to-date and correct, no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability or availability with respect to the website or the information, products, services, or related graphics contained on the website are made for any purpose. Any reliance placed on such information is, therefore, strictly at user's own risk.

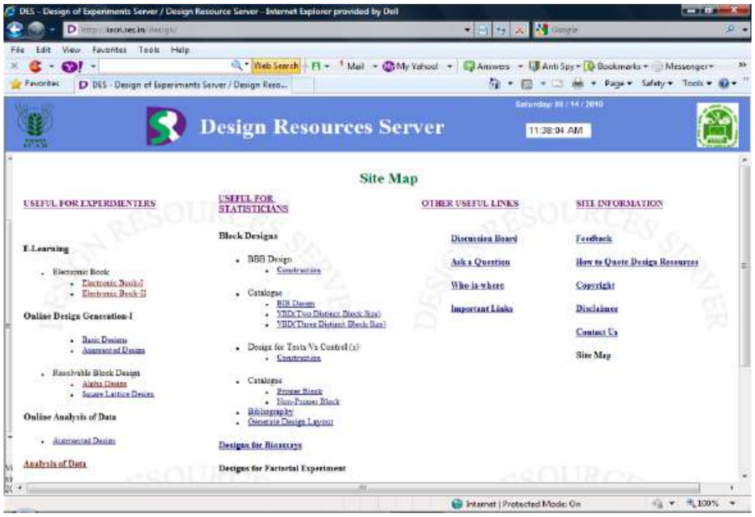
In no event will "IASRI" be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from loss of data or profits arising out of or in connection with the use of this website.

Through this website users are able to link to other websites which are not under the control of "IASRI". The inclusion of any links does not necessarily imply a recommendation or endorsement the views expressed within them.

Every effort is made to keep the website running smoothly. However, "IASRI" takes no responsibility for and will not be liable for the website being temporarily unavailable due to technical issues beyond our control.

5.5 Site Map

This link gives a map of the various links available on the server. A user can access any of the links through this map also. A snap shot of the site map is given below:


DES - Design of Experiments Server / Design Resource Server - Internet Explorer provided by Dell
http://iasri.res.in/design/
File Edit View Favorites Tools Help
Web Search
Mail
My Yahoo!
Answers
Anti Spy
Bookmarks
Messenger
Favorites
DES - Design of Experiments Server / Design Reso...
Page
Safety
Tools
Design Resources Server
Saturday 08 / 14 / 2010
11:38:04 AM
Internet | Protected Mode: On
100%

Site Map

USEFUL FOR EXPERIMENTERS

E-Learning

- Electronic Book
  - Electronic Book-I
  - Electronic Book-II

Online Design Generation-I

- Basic Designs
- Augmented Designs
- Resolvable Block Designs
  - Alpha Designs
  - Square Lattice Designs

Online Analysis of Data

- Augmented Designs

Analysis of Data

USEFUL FOR STATISTICIANS

Block Designs

- BBB Design
  - Construction
- Catalogue
  - BIB Designs
  - VBD(Two Distinct Block Size)
  - VBD(Three Distinct Block Size)
- Design for Tests Vs Control (s)
  - Construction
- Catalogue
  - Proper Block
  - Non-Proper Block
- Bibliography
- Generate Design Layout

Designs for Bioassays

Designs for Factorial Experiment

OTHER USEFUL LINKS

Discussion Board

Ask a Question

Who is where

Important Links

SITE INFORMATION

Feedback

How to Quote Design Resources

Copyright

Disclaimer

Contact Us

Site Map

6. Some Information on the Usage of the Server

- Design Resources Server is a copyright of IASRI (ICAR). The Server was registered under Google Analytics on May 26, 2008. For the period May 26- October 31, 2011, it has been used through 1102 cities in 113 countries spread over 6 continents. The average time on the page is 2.59 minutes.
- External links of the server are also available at:
  - http://en.wikipedia.org/wiki/Design_of_experiments
  - http://en.wikipedia.org/wiki/Hadamard_matrix
- The server has been cited at:
  - https://dspace.ist.utl.pt/bitstream/2295/145675/1/licao_21.pdf for lecture presentation on Unitary operators.
  - Chiarandini, Marco (2008). DM811-Heuristics for Combinatorial Optimization. Laboratory Assignment, Fall 2008. Department of Mathematics and Computer Science, University of Southern Denmark, Odense.
  - http://support.sas.com/techsup/technote/ts723.html

  - Warren F. Kuhfeld. Orthogonal Arrays. Analytics Division SAS, Document No. 273 (http:// support.sas.com/techsup/technote/ts723.html).
  - Electronic text material in "New and Restructured Post-Graduate Curricula & Syllabi on Statistical Sciences (Statistics/Agricultural Statistics; Bio-Statistics, Computer Application) of Education Division, Indian Council of Agricultural Research, New Delhi, 2008.
  - Jingbo Gao, Xu Zhu, Nandi, A.K. (2009). Nonredundant precoding and PARR reduction in MIMO OFDM systems with ICA based blind equalization. IEEE transactions on Wireless Communications, 8(6), 3038-3049.
- Server is also linked at
  - ICARDA Intranet: Biometric Services
  - CG Online learning resources- http://learning.cgiar.org/moodle/Experimental Designs and Data Analysis

7. Future Directions

The Design Resources Server created and being strengthened at IASRI aims to culminate into an expert system on design of experiments. To achieve this end, the materials available on various links need to be strengthened dynamically. Besides this, the following additions need to be made to the server in the near future:

- Online generation of
  - balanced incomplete block designs, binary balanced block designs and partially balanced incomplete block designs
  - block designs with nested factors
  - designs for crop sequence experiments
  - efficient designs for correlated error structures
  - online generation of row-column designs
  - designs for factorial experiments; fractional factorial plans
- designs for microarray experiments
- designs for computer experiments
- designs for fitting response surfaces; designs for experiments with mixtures
- split and strip plot designs
- field book of all the designs generated
- labels generation for preparing seed packets
- online analysis of data

The success of the server lies in the hands of users. It is requested that the scientists in NARS use this server rigorously and send their comments for further improvements to Dr. Rajender Parsad (Rajender.Parsad@icar.gov.in). The comments/ suggestions would be helpful in making this server more meaningful and useful.