BIOVERSITY TECHNICAL BULLETIN NO. 14

Bioversity International

Data Analysis Manual
for Coconut Researchers

P. N. Mathur, K. Muralidharan, V. A. Parthasarathy, P. Batugal and F. Bonnot

COGENT

Bioversity Technical Bulletins are published by Bioversity International with the intention of putting forward definitive recommendations for techniques in genetic resources. They are specifically aimed at National Programmes.

Previous titles in this series:

A protocol to determine seed storage behaviour
T.D. Hong and R.H. Ellis
IPGRI Technical Bulletin No. 1, 1996.

Molecular tools in plant genetic resources conservation: a guide to the technologies
A. Karp, S. Kresovich, K.V. Bhat, W.G. Ayad and T. Hodgkin
IPGRI Technical Bulletin No. 2, 1997.

Core collections of plant genetic resources
Th.J.L. van Hintum, A.H.D. Brown, C. Spillane and T. Hodgkin
IPGRI Technical Bulletin No. 3, 2000.

Design and analysis of evaluation trials of genetic resources collections
Statistical Services Centre and University of Reading
IPGRI Technical Bulletin No. 4, 2001.

Accession management: combining or splitting accessions as a tool to improve germplasm management efficiency
N.R. Sackville Hamilton, J.M.M. Engels, Th.J.L. van Hintum, B. Koo and M. Smale
IPGRI Technical Bulletin No. 5, 2002.

Forest tree seed health
J.R. Sutherland, M. Diekmann and P. Berjak
IPGRI Technical Bulletin No. 6, 2002.

***In vitro* collecting techniques for germplasm conservation**
V.C. Pence, J.A. Sandoval, V.M. Villalobos A. and F. Engelmann
IPGRI Technical Bulletin No. 7, 2002.

Análisis Estadístico de datos de caracterización morfológica
T.L. Franco y R. Hidalgo
IPGRI Technical Bulletin No. 8, 2002.

A methodological model for ecogeographic surveys of crops
L. Guarino, N. Maxted and E.A. Chiwona
IPGRI Technical Bulletin No. 9, 2005.

Molecular markers for genebank management
D. Spooner, R. van Treuren and M.C. de Vicente
IPGRI Technical Bulletin No. 10, 2005.

***In situ* conservation of wild plant species a critical global review of good practices**
V.H. Heywood and M.E. Dulloo
Bioversity Technical Bulletin No. 11, 2006

Crop genetic diversity to reduce pests and diseases on-farm. Participatory diagnosis guidelines. Version 1.
D.I. Jarvis and D.M. Campilan
Bioversity Technical Bulletin No. 12, 2007

Developing crop descriptor lists: guidelines for developers
Bioversity Technical Bulletin No. 13, 2007

Copies can be obtained in PDF format from Bioversity's Web site (www.bioversityinternational.org) or in printed format by sending a request to bioversity-publications@cgiar.org.

Data Analysis Manual for Coconut Researchers

P.N. Mathur[1], K. Muralidharan[2], V.A. Parthasarathy[3], P. Batugal[4] and F. Bonnot[5]

[1] Bioversity International, Sub-regional Office for South Asia, NASC complex, Pusa campus, New Delhi 110012, India

[2] Central Plantation Crops Research Institute, Kasaragod 671124, Kerala, India

[3] Indian Institute of Spices Research, Post Box No. 1701 Marikunnu PO, Kozhikode 673012, Kerala, India

[4] Bioversity International, Regional Office for Asia the Pacific and Oceania, Serdang, Selangor DE, Malaysia

[5] Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) Avenue Agropolis, 34398 Montpellier Cedex 5, France

Bioversity International is an independent international scientific organization that seeks to improve the well-being of present and future generations of people by enhancing conservation and the deployment of agricultural biodiversity on farms and in forests. It is one of 15 centres supported by the Consultative Group on International Agricultural Research (CGIAR), an association of public and private members who support efforts to mobilize cutting-edge science to reduce hunger and poverty, improve human nutrition and health, and protect the environment. Bioversity has its headquarters in Maccarese, near Rome, Italy, with offices in more than 20 other countries worldwide. The Institute operates through four programmes: Diversity for Livelihoods, Understanding and Managing Biodiversity, Global Partnerships, and Commodities for Livelihoods.

The international status of Bioversity is conferred under an Establishment Agreement which, by January 2006, had been signed by the Governments of Algeria, Australia, Belgium, Benin, Bolivia, Brazil, Burkina Faso, Cameroon, Chile, China, Congo, Costa Rica, Côte d'Ivoire, Cyprus, Czech Republic, Denmark, Ecuador, Egypt, Greece, Guinea, Hungary, India, Indonesia, Iran, Israel, Italy, Jordan, Kenya, Malaysia, Mali, Mauritania, Morocco, Norway, Pakistan, Panama, Peru, Poland, Portugal, Romania, Russia, Senegal, Slovakia, Sudan, Switzerland, Syria, Tunisia, Turkey, Uganda and Ukraine.

Financial support for Bioversity's research is provided by more than 150 donors, including governments, private foundations and international organizations. For details of donors and research activities please see Bioversity's Annual Reports, which are available in printed form on request from bioversity-publications@cgiar.org or from Bioversity's Web site (www.bioversityinternational. org).

Citation: Mathur, P.N., K. Muralidharan, V.A. Parthasarathy, P. Batugal and F. Bonnot. 2008. Data Analysis Manual for Coconut Researchers. Bioversity Technical Bulletin No. 14. Bioversity International, Rome, Italy.

ISBN: 978-92-9043-736-9

Bioversity International
Via dei Tre Denari, 472/a
00057 Maccarese
Rome, Italy

Introduction to the Series

The Technical Bulletin series is targeted at scientists and technicians managing genetic resources collections. Each title will aim to provide guidance on choices while implementing conservation techniques and procedures and the experimentation required to adapt these to local operating conditions and target species. Techniques are discussed and, where relevant, options presented and suggestions made for experiments. The Technical Bulletins are authored by scientists working in the genetic resources area. Bioversity welcomes suggestions of topics for future volumes. In addition, Bioversity would encourage, and is prepared to support, the exchange of research findings obtained at the various genebanks and laboratories.

Table of Contents

List of Tables

List of Figures

Foreword

In the last 50 years, coconut breeders have been conducting intensive research on the coconut in order to improve total farm productivity and enhance farmers' income. A major part of this effort is the identification and characterization of coconut diversity and the development of varieties and hybrids which are high-yielding, resistant to biotic and abiotic stresses, and possessing important traits for producing high-value products.

A major problem of this effort is the long gestation period of these research activities. This is partly due to the perennial nature of the coconut which takes 4-12 years before resulting varieties and hybrids start to produce fruits. This problem is further complicated by the lack of standardized methods of coconut breeding which has prevented coconut researchers from comparing results. An equally important factor is the fast turnover of coconut breeders as some of them retire, occupy administrative posts, or shift to other professional activities. This requires that coconut research institutes conduct training of young coconut researchers on coconut genetic resources and improvement. Unfortunately, many countries do not have resources or are incapable of conducting this training.

To address the above problems, the International Coconut Genetic Resources Network (COGENT) published in 1994 the "Manual on standardized techniques in coconut breeding" (STANTECH). While the manual has been useful, it does not contain enough details to enable coconut researchers to more effectively conduct experiments and analyze and interpret results. To complement this publication, COGENT is publishing this "Data Analysis Manual for Coconut Researchers". The manual does not only provide basic statistical concepts but also details of experimental designs to use in agronomic and multilocation trials and methods for analysis of multilocation trials data and determination of genetic distance. The manual uses actual coconut data in demonstrating how data can be analyzed and interpreted. Using this tutorial-type manual, coconut researchers can practise computations by themselves using the examples and then use the same procedure to analyze their own data.

I would like to thank the Central Plantation Crops Research Institute (CPCRI), and the Indian Institute of Spices Research (IISR) of the Indian Council of Agricultural Research (ICAR), and The

French Agricultural Research Centre for International Development (CIRAD) for allowing their staff to help develop this manual. I also thank the International Fund for Agricultural Development (IFAD) for supporting its publication.

Richard Markham
Director
Commodities for Livelihood Programme
Bioversity International

Chapter 1: Basic statistical concepts

The science of statistics deals with the methods of collection, presentation, analysis and interpretation of data. Applications of such methods are not necessary when all the units under observation indicate no variation or follow a rigid mathematical law. In such cases, observation on a few units would provide all the needed information about the population or of the units. However, in biological science, such uniformity is rare and it is necessary to take recourse to the statistical science to secure sound methods of collecting data and appropriate techniques for analyzing the data to derive reliable conclusions. In this context, statistics may well be defined as the science of the study of variations. Consider the problem with the objective of characterization of coconut tall population for nuts per bunch in a garden. The researcher can either count nuts per bunch from all the coconut trees of tall population in that garden or count only from few selected trees. It is obvious that the number of nuts vary between bunches of a tree and also between trees. To characterize the population one has to summarize the observations made on the individuals (either for whole of the population or a part of it). By applying appropriate statistical techniques, observation on a few units would provide all the information needed about the population or the totality of units. The quality of information generated by applying these methods and the inferences thus drawn largely depends on the appropriateness of procedures used.

There are two categories of statistics: descriptive statistics and inferential statistics. A set of observations obtained from individuals, comprising the *universe*, can be processed and summarized using textual, tabular and graphical methods of data presentation. Frequency distribution tables (FDT), bar graphs, or polygons help to visualize the distribution of the observations. Numerical descriptive measures which include measures of central tendency and measures of dispersion are effective ways to summarize a voluminous set of data into a significantly smaller number of values. On the other hand, inferential statistics involves estimation and tests of hypotheses. Estimation includes different methods of sampling, choice of estimators, and tests of hypotheses about parameters of a single population or those of two or more populations. To allow valid comparisons, experiments should follow the basic principles of experimental designs and must meet the assumptions of the statistical procedures. Specific statistical procedures such as genetic distance estimation genetic diversity assessment and genotype x environment interaction are especially applicable for coconut genetic resources evaluation and multilocation trials. Multivariate procedures such as cluster analysis and principal component analysis, while also descriptive, summarizing observations on many characters, can be used to test inferences about the populations as well. For better appreciation, each of these methods will be discussed in detail with real data from coconut experiments. The data used for most of the examples in this manual are actual

research data on coconut generated at the Central Plantation Crops Research Institute (CPCRI), Kasaragod, India. Almost all the analyses described in this manual have been described with some suitable examples and step-by-step procedures. It is therefore expected that this publication will help the coconut scientific community to better plan and manage their experiments and, analyze and interpret their experimental data.

Definition of terms

Meanings of some of the basic statistical terms which will be used throughout this manual and which are sometimes confusing are defined below:

Universe

A *universe* consists of the totality of units or individuals under study. It may be a variety or an open-pollinated population of coconut palms. The individual palms are the individual units comprising the universe.

Variable

A characteristic that may vary from unit to unit is called a *variable*. In a set of coconut palms, stem height is a variable and so are stem diameter, leaf morphology, length of central axis, fruit and nut shapes.

Population

A *population*, statistically speaking, consists of all possible values of the variable or characteristic of interest. Hence $E = \{x : 12m \leq x \leq 30m\}$ is a population induced by stem height (denoted as x) on the universe under study and because of this they are interchangeably used.

Qualitative variable

A variable is said to be *qualitative* if its values reflects quality, attribute or categories. The colour of seedling, shape of crown, leaf spiral direction, presence or absence of trichomes, etc. are examples of some quality characters in coconut.

Quantitative variable

A variable is said to be *quantitative* if its values reflects magnitude or amount of an attribute like measurements or counts. Quantitative variables may either be *continuous* or *discrete*. Stem height, days to 50% flowering, leaf number, width of leaf scar, fruit weight, copra yield, etc. are some examples of quantitative characters in coconut.

Continuous variable

A quantitative variable that take values within an interval characterized by an infinite set of possible values is said to be continuous. Crop yield, plant height, plant weight, temperature and volume are examples of continuous variables.

Discrete variable

A quantitative variable that take only certain specific values usually integral values within a given range is said to be discrete. The number of nuts in a bunch, the number of fruit bearing trees in a plantation and the number of leaf scars are examples of discrete variables. Two forms of discrete data may be recognized viz., attributes and counts. The first form, which can be related to qualitative type of data, classifies individuals as having or not having a particular attribute or more commonly, describes a group of individuals by the proportion or percentage of individuals possessing a particular attribute. Some examples are proportion of coconut trees infected by leaf blight, seedling survival rate (expressed as per cent), etc. In the second form, the individual is described by a numerical count that cannot be expressed as proportion. Some of the examples for coconut are number of first split leaf, number of leaflets, number of spikelets with female flowers and number of spikelets without female flowers.

A distinction is made between continuous and discrete variables because the two types of data may require different statistical analysis. Most of the sampling methods and computational procedures described are applicable for continuous variables. The analytical procedures for discrete variables are generally more complex. By increasing the number of values that a discrete variable can assume, however, it is often possible to handle such data by the continuous variable methods.

Sample

A sample is a subset of the universe or a population about which we wish to draw information. Using the information obtained from the sample, statistical inferences about the population are made.

Parameter

Parameters are numerical measures that describe the population and distinguish it from other populations. Once the parameters are known they completely specify the mathematical form of the distribution of the observations. The exact values of these parameters are obtained if and only if all the members of the universe are observed.

Statistic

By observing a subset of the population, the population parameter is estimated by some rule or estimator which is called a statistic. Being a characteristic of the sample, values of the statistic may vary from sample to sample.

Statistical inference

The statements that can be made about the parameters of a population or the form of its distribution based on the information contained in the samples. Inferences, however, can only be made when samples used are probability samples.

Point estimate

It is a single number stated as an estimate of some quantitative property of the population. For example, the average stem height of a coconut population is 25 m, 3.5% of coconut trees were infected with leaf blight, or 45 plants in a coconut germplasm accessions have egg-shaped fruit polar section shape.

Interval estimate

It is a statement that a population parameter has a value lying between two specified limits. For example, the average stem height of a coconut population is between 23 and 27 m, coconut germplasm collection from any country may have 15 to 20% accessions with erect stem.

Confidence interval

It is one type of interval estimate. In repeated sampling, a known proportion of the intervals computed by this method would include the population parameter. For example, the 95% confidence interval of average stem length at 11 leaf scars in a coconut population is between 5.2 and 12.1 dm.

Random sampling

Random sampling in its simplest form is a method of drawing a sample such that each member of the population has an equal chance of being included in the sample.

Sampled population

It is the population to which statistical inferences from the sample apply. The process by which samples are obtained gives every member of the population a known chance of being represented. In practice, the sampled population is sometimes hypothetical rather than real, because the only available data may not have been drawn at random from a known population.

Target population

The target population is the aggregate about which the investigator is trying to make inferences from the sample. Although this term is not in common use, it is sometimes helpful in focusing attention on differences between the population actually sampled and the population that we are attempting to study.

Frequency distribution

In a frequency distribution, the values in the sample are grouped into a limited number of classes. A table is made showing the class boundaries and the frequencies (number of members of the population/sample) in each class. The purpose is to obtain a compact summary of the data.

Null hypothesis

A null hypothesis is a specific assumption or belief about a population that is being tested for its validity by means of appropriate data analysis of the samples or treatments.

Test of significance

A test of significance is, in general term, a calculation by which the sample results are used to throw light on the truth of a null hypothesis. It measures the extent to which the sample departs from the null hypothesis in some relevant aspect. If the value of the test criterion falls beyond certain limits into a *region of rejection*, the departure is said to be *statistically significant*. Tests of the level of significance have a known value, most commonly 0.05 (significant) or 0.01 (highly significant).

Critical difference

It is the value, equal to or greater than which the difference between two treatment effects is significant.

Degrees of freedom

The number of independent comparisons that can be made between the members of a sample (e.g., subjects, test items, trials, conditions, etc.). The number of degrees of freedom is one less than the number of variates in the sample concerned.

Primary data

Data collected by an investigator or researcher for the purpose of the current investigation are called primary data. Usually, the primary data are collected from field or laboratory trials or from observations on a sample of individuals/units of a population.

Secondary data

Data taken from existing records maintained by various institutions for some other purpose. The area and production statistics of coconut documented by government agencies and used by researchers are examples of secondary data.

Chapter 2: Sampling methods

When the population size is large, it is not possible to make observation on all the units constituting that population. In such situations information on the population can be drawn by observing only a few units. The principles employed for selection of a sample of units from a population are described in this chapter, and subsequently the sampling strategy for coconut germplasm collecting.

Reasons for sampling

Consider the problem of estimating the total coconut production for a given locality in a given cropping season. Clearly, it is impractical to harvest and weigh the produce from all the fields growing coconut in the locality which constitute the population under study. It is obvious that in cases, where the population is too large, the investigator has to obtain information about the population from only a part thereof. The process of selecting a part of the population to represent the entire population under study is known as *sampling* and the part selected is known as *sample*. The manner by which we draw the sample hardly matters provided the population is homogeneous with respect to the character of interest. However, when the units of the population vary considerably, the method by which we draw the sample plays a critical role.

Two types of samples arise based on how they were obtained. Non-probability samples constitute those samples in which the elements were selected on purpose while probability samples are those that were obtained using some random mechanism. Between the two, it is the probability samples that are of interest as only with these samples can we make valid statistical inferences about the population. We are more interested in probability sampling as it provides valid estimates of error, based on probability theory, and objective conclusions could be drawn. There would be some exception to random sampling and we will discuss this at appropriate places (for example, selection of elite coconuts from a population).

Some of the reasons why we sample include: 1) reduced cost, 2) greater speed in making the results available, 3) greater scope, 4) greater accuracy and 5) necessity. Because of limited resources in the form of money, trained personnel or some specialized equipment, it is not feasible to observe every member of the population. With only a few units observed, more information can be obtained with greater accuracy on the units.

Principal steps in sampling

In any survey activity, the objectives should be clearly stated. Details in planning can easily obscure the goals of the survey when they are not clear and defined. Defining the sampling units and population is fundamental to any sampling process. The list of sampling units that divide the population into non-overlapping

parts is called the *sampling frame.* Other considerations includes: 1) degree of precision desired, 2) availability of resources, 3) selection of the sample and 4) organization of the fieldwork. The sample size for specified precision needs special consideration as it should be neither too large nor too small. The sample size required for measuring the population parameters with statistically acceptable precision at a given cost could be worked out by using suitable formulae (For more details refer to Cochran 1946 and 1984).

Methods of probability sampling

Simple random sampling

Simple random sampling is a method of selection of '*n*' individual units out of a population of 'N' units, such that each unit of the population unit has an equal chance of being included in the sample. In practice, a table of random numbers is used for selection of units and sampling is done without replacement. The application of simple random sampling presumes the population under study to be divisible into a number of distinct identifiable units from which selection can be done.

The practical difficulty in the way of random selection has been largely overcome by the use of published tables of random numbers. They usually consist of columns of 1, 2, 3 or 4 digit numbers randomly drawn and tested for randomness. To use these tables, the units constituting the populations are to be numbered. The size of the population determines the random number table of 1, 2, 3, or 4 digits to be used. If the population size is less than 10, we use one digit random numbers. If it is less than 100, 1000, or 10000, we use 2, 3, or 4 digit random number tables, respectively. These tables could also be used in the case of larger population sizes. Starting any where in these tables, customarily one moves down the column then to the next columns to choose the numbers in the populations. For the purpose of illustration, the set of random numbers arranged in Table 2.1 will be used.

Illustration

Suppose we have to choose a sample of five coconut palms from a population size of 80. In this case, we make use of two digit random numbers as the population size is less than 100. Starting at random in the table of random numbers (Table 2.1) and moving down, we select random numbers say, 21 followed by 87, 91, 73, 96, 27, 23, 23, 18, etc. The coconut palms corresponding to the numbers drawn are included in the sample. Since the first random number is 21, palm no. 21 in the population is selected. Selected numbers greater than 80 or the population size is therefore rejected since no palms will be numbered as such. The third unit selected corresponding to palm no. 27, followed by palm no. 23. Note that the next random number is again 23. If the sampling is being done *with replacement,* the unit number 23 is again included in the sample. For *sampling without replacement*, this number is rejected. The process is continued until the required number of units has been selected. In this example, for sampling without replacement, the sample

will consist of units 21, 73, 27, 23 and 18. However, in the case of sampling with replacement, the selected unit numbers will be 21, 73, 27, 23 and 23. Other methods of using tables of random numbers for various purposes can be obtained in Fisher and Yates (1963).

The mean of the observations on the sampled units is called the *sample mean*. This is used to gain knowledge on the mean of the population. In other words, the sample mean is the estimator of the population mean. As might be expected, the various sample means differ among themselves. The standard deviation of all possible sample means is commonly referred as *standard error*. The square of the standard error is called the *sampling variance of the mean*. It is important to note that standard error involves a factor $f = (N-n)/N$, the finite population correction. Denoting the 'sample variance' as s^2 (obtained by dividing the sum of squares by n-1), the estimate of standard error is given by $s\sqrt{(f/n)}$.

Sometimes, it is not possible to recognize and number the individual units within the population. For example, if we wish to select a sample of random nuts from a coconut plantation for taking observations, it will not be possible to recognize individual nuts and number them. Under such conditions, alternative procedures of random selection are available. Instead of numbering the nuts, we could number the palms first based on random selection and thereafter select the nut(s) at random from within the selected palms for inclusion in the sample. This procedure of selecting a random sample in successive stages is called *sub-sampling or multi-stage random sampling* and is discussed later in this chapter. This device is extensively used in sampling due to ease in selection and economy of labour.

Stratified sampling

It is obvious that a reduction in 'sample variance' will lead to estimates with less standard error. Stratified random sampling is a method that takes advantage of known information about the population to reduce sample variance. In stratified random sampling, the N units in the population are grouped into L sub-populations or *strata* on the basis of similarity with respect to some characteristic. The groups or sub-populations are of sizes N_1, N_2, ... N_L such that $N_1 + N_2 + ... N_L = N$. A random sample from each stratum is obtained to give an estimate of the stratum mean. The estimates of the different strata are then combined to give an estimate of the population mean.

In sampling large coconut populations, we stratify the palms into the major palm types, make separate sample estimates for each type, and then combine to get an estimate for the entire population. If the variation among units within a stratum is less than the variation among units of different strata then the combined estimates will be more precise. This sampling procedure is widely used due to its convenience, especially when the stratification is adopted according to geographical contiguity or administrative classification such as state, province/district, block, village, etc. Another useful basis for stratification is agroclimatic zones/classification.

Table 2.1. Section of 10000 random digits

8	8	7	5	8	6	6	6	0	5
3	5	6	6	1	4	2	8	3	2
2	6	3	3	5	0	3	7	7	1
6	0	8	2	6	7	4	7	1	8
9	5	0	4	4	9	9	8	9	6
8	3	7	4	6	4	7	6	9	4
2	7	9	9	8	4	2	5	6	2
8	2	6	8	5	3	2	3	2	3
1	8	3	8	6	1	3	8	6	2
2	1	7	1	7	1	3	1	4	1
1	8	4	4	6	8	3	0	5	2
6	6	0	2	7	7	5	1	7	7
5	1	4	2	0	9	6	7	7	9
2	7	0	4	5	6	2	6	2	6
1	3	0	9	4	1	7	7	2	5
9	2	3	8	2	6	2	5	1	8
1	6	2	1	5	5	0	8	0	9
0	9	3	4	2	1	4	5	2	8
3	8	1	4	8	7	9	0	0	1
2	3	6	8	9	1	9	9	9	7
2	5	4	0	7	3	7	7	2	6
2	5	3	4	9	6	9	4	5	6
0	2	3	2	2	7	7	4	9	1
1	5	0	7	2	3	3	2	6	1
2	7	0	0	2	3	1	0	3	6
6	6	1	8	1	8	3	3	1	6
0	9	7	7	9	0	1	8	2	2
1	0	7	9	1	0	7	7	0	6
7	4	8	3	3	5	5	7	6	7
1	7	5	8	3	2	4	0	3	

Stratified random sampling offers two primary advantages over simple random sampling. First, it provides separate estimates of the mean and variance of each stratum (e.g. irrigated and rainfed coconut plantation areas). Second, for a given sampling intensity, it often gives more precise estimates of the population parameters than would a simple random sample of the same size. For the latter advantage, however, it is necessary that the strata be set up in such a way that the variability among unit values within a stratum is less than the variability among units from different strata.

Some drawbacks of stratified sampling however are: 1) each unit in the population has to be assigned to one and only one stratum, 2) the size of each stratum has to be known, and 3) sample has to be taken from each stratum.

The most common constraint in using stratified random sampling is the lack of information on the stratum sizes. If the sampling fractions are small in each

stratum, it is not necessary to know the exact stratum size. The population means and standard errors can be computed from the relative sizes.

Cluster sampling

In some situations, the sampling units exist in naturally formed groups (clusters). It will then be convenient and cost effective to sample the clusters. All the units in the selected cluster should be included in the sample.

Systematic sampling

The sample in this method is obtained by selecting units at fixed intervals in the population. Estimation of error and possible bias arising from unrecognized trends or cycles in the population are the drawbacks of this method.

To select sample from an area, systematic sampling in two dimensions may be used. Here, systematic sample points are fixed for both directions (East to West and North to South) separately and then form a grid by repeating the sample points of one direction through every sample point selected in the other direction. The intersecting points form the sample. Instead of the 'square grid' pattern of sample selection, an 'unaligned' sample can also be selected.

Multi-stage sampling

If each unit in a sample can be sub-divided into a number of smaller units or elements, further sampling can be done from each of the selected first stage units. The sampling design is then termed two-stage sampling. The sub-divisions of the first stage units may sometimes further be divided and sampled as third stage units and so on. Multi-stage sampling design is popular in crop surveys because of its flexibility in the selection of units.

Sampling strategy for coconut germplasm collecting

Knowledge on the extent of coconut growing areas, percentage area under the crop and the degree of environmental diversity are used to stratify the location identified for germplasm expedition. However, a basic problem in coconut germplasm collecting arises from the fact that the crop has a cosmopolitan distribution with an uncertain centre of origin.

Guided by the sampling methods described above, the coarse and fine grid sampling strategies (Santos *et al.* 1996) outlined below would ensure that the areas to be sampled are carefully scrutinized and that a minimum area is skipped. Since there is no existing information on natural gene flow between coconut populations (apart from the studies conducted on polyphenols, electrophoresis, and more recently, DNA patterns among coconut varieties obtained from various origins, all of which leads to variable and inconclusive results), the coarse grid sampling strategy as described in a practical course initially organized by the Bioversity International (Formally known as IPGRI and IBPGR) in 1978 and 1979 in Bogor, Indonesia, has been tried leading to a potential systematic coverage of the coconut areas in the Philippines (Santos, 1987) and Malaysia (Jamadon, 1987). When combined

with fine-grid and biased sampling method, this strategy ensures that no important diversity is missed, and that the widest possible array of conditions is covered. Another feature of this strategy is that surveys can be suspended and resumed as necessary by the explorer/collector because areas are pre-identified.

Coarse grid sampling method

The basic requirements for coarse grid sampling method are:

1. Knowledge of the extent of coconut-growing areas and distribution, and
2. Knowledge of the degree of environmental diversity in the areas where the crop is grown.

The coarse grid sampling procedure is demonstrated as implemented by the Breeding and Genetics Division of the Philippine Coconut Authority (PCA). The procedure is as follows:

1. A suitable sized map of the Philippines is obtained (Scale: 1:1,000,000) and grids of approximately 40 x 40 km were marked (Fig. 2.1), following latitude and longitude divisions/degrees.
2. All grids which included coconut-growing areas are then identified and according to relative size or hectarage, the number of sampling sites is determined per grid. (See item 6 below).
3. The leader of the survey team then makes travel arrangements with the local PCA Regional and Provincial Offices so that all the needed equipment, materials and supplies are prepared in advance.
4. The two-person team (e.g. the leader - a technical staff of the division and a climber/technician) set out to travel to the pre-identified sites.
5. Local laborers are hired to assist in the fruit component analysis (FCA) and vegetative measurements according to the minimum list of descriptors, and passport data.
6. Each grid is surveyed using 5 to 6 sampling sites (SS) with intervals between SS determined according to any noticeable changes in ecological conditions, e.g. coastal to upland.
7. A local guide (normally the coconut development officer) assists the team in explaining the objectives of the survey to the farmers/owners and to the local population.

Biased sampling/outright collecting

1. If the technical person encounters something new, which he/she feels is not represented in the genebank, the supervisor is notified and he/she will determine whether the prospective population is unique or not. If it is, and no important disease or pest is evident, then the decision is made to collect 200 nuts of the population or variety. These will be sent to the research centre for conservation in the genebank.

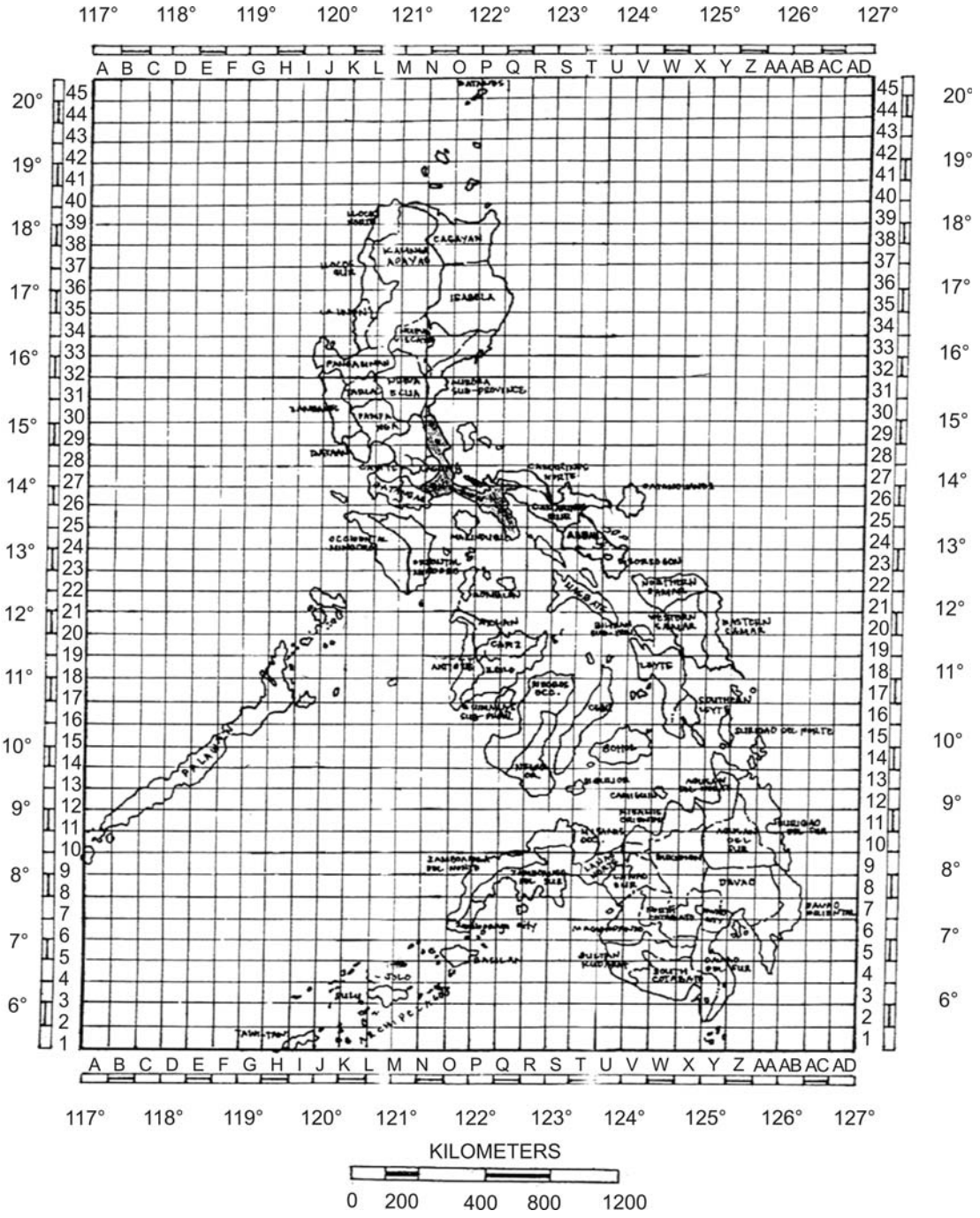


Figure 2.1. Map of the Philippines used for coarse grid sampling (*Source:* Santos 1987).

Note: *Since good information on the distribution of diseases in the country (the Philippines) exists, the above step is possible. But, it is emphasized that this is only done when one is sure that the risk is negligible. Otherwise, collecting should be made only after ascertaining that the transfer of the material is safe and sound.*

2. After surveying 4 to 5 grids equivalent to 20-30 sampling sites, data are processed which includes estimation of genetic distances of the surveyed populations. Information on genetic diversity of the surveyed populations are obtained using cluster analysis.
3. The survey is always coordinated with local officials and extension workers because, depending on the result of the genetic partitioning, field collecting of seed nuts from the identified farms are coordinated with the local guides.

Collecting seed nuts

Depending on the results of statistical analysis, nuts are sampled in such a way that the widest array of genotypes/phenotypes in the identified population(s) is covered. If a certain grid is noted to be different in ecological condition from that of the other grids but without any obvious significant differences noted between sites, the same number of samples is collected for better coverage. The collection is identified in the genebank as a unique accession to be assigned its own accession number (Acc No.___) noting the grid number (Grid No. __) and the sampling site number (SS No.___). The collection is planted at random within the designated block with each palm properly identified as to its exact origin, e.g. Grid # A14, SS # 3.

References

Cochran, W.G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals of Mathematical Statistics, 17: 164-177.

Cochran, W.G. 1984. Sampling techniques, 2nd edition. Wiley and Sons, New York, 413p.

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research. 6th edition. Longman, Edinburgh. 146p.

Jamadon, B. 1987. Coconut genetic resources activities in Malaysia. Report of the First Meeting, UNDP/FAO Project RAS/80/032/G/01/02. Working Group on Genetic improvement. IBPGR. Pp. 32-45.

Santos, G.A. 1987. Activities of coconut genetic resources collection and conservation in the Philippines. Report of the First Meeting. UNDP/FAO Project RAS/80/032/g/01/12. Working group on Genetic Improvement. IBPGR. Pp. 56-72.

Santos, G.A., Batugal, P.A., Othman, A., Baudouin, L. and Labouisse, J.P. 1996. Manual on standardized research techniques in coconut breeding. IPGRI/COGENT, Serdang, Malaysia. 46p.

Further Reading

Rao, Poduri S.R.S. 2000. Sampling Methodologies with Applications. Chapman and Hall/CRC, New York.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. 1977. Foundations of Inference in Survey Sampling. John Wiley and Sons, New York.

Lehtonen, R. and Pahkinen, E.J. 1995. Practical Methods for Design and Analysis of Complex Surveys. John Wiley and Sons, New York.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. 1984. Sampling Theory with Applications. Iowa State University Press and Indian Society of Agricultural Statistics, New Delhi.

Chapter 3: Frequency distribution of observations

As a standard procedure, both for qualitative or quantitative data, the researchers keep records of their observations, corresponding to each individual under study. Faced with voluminous data, researchers need to come up with methods that can summarize the information contained in these datasets into fewer but concise measures with very little loss of information. There are several ways of summarizing and presenting data. Data can be summarized using textual, graphical and tabular methods. Numerical descriptive measures which also described the population in terms of one or more quantitative measures are frequency distribution and is a first step towards summarizing data. The mathematical descriptions of several types of frequency distribution are known. The form of distribution pattern is determined by the values of the constants of its mathematical function. These constants are referred to as the parameters of the distribution. Further, summarization of data is possible in terms of parameters and can be used for comparison. This Chapter deals with certain useful parameters of frequency distributions, in particular, that of normal distribution.

Frequency distribution

One information that is of prime interest is how certain variables are distributed in the populations under study. A most useful procedure for obtaining and presenting this information is through the use of frequency distribution tables. Frequency distribution tables group the observations into non-overlapping classes and present both the frequencies and relative frequencies of the different classes. This way, one gains an idea of which classes dominate or which classes are present in the population. The importance of frequency distribution tables lies in the fact that it can be used for both qualitative and quantitative characters.

Frequency distribution of qualitative data

In case of qualitative data, such as fruit shape, individual plants would be classified according to the fruit shape and the corresponding number for each class will be counted.

Example

In a population of 314 coconut palms, the shape of the nut in each palm was classified as either round, egg-shaped, pear-shaped and elliptic. The four different shapes are the 'classes' and the corresponding number of palms having nuts with a specific shape is the frequency of that class. The frequency distribution thus obtained is shown in Table 3.1. The relative frequencies expressed as percentages will have the advantage of easy interpretation. For example, examining the Table 3.1, it is easy to state that more than half of the palms have round nuts.

Frequency distribution of quantitative data

The frequency distribution of quantitative characters, on the other hand, necessitates an arbitrary classification of the observations under study. For instance, to measure the yield of nuts per palm from a coconut plantation, the researcher will obtain measurements on a large number of palms from a coconut field under study. These measurements may take any value within a certain range. After recording the data, the researcher's first task will be to classify these data with the object of reducing them to a form in which they can be conveniently handled. The classification of such data involves the partitioning of the range of values into a number of non-overlapping but contiguous class intervals and recording the number of individual observations falling in each class.

Table 3.1. Frequency distribution of fruit shape in coconut population

Class	Frequency	Relative Frequency
Round	169	53.82
Egg-shaped	61	19.43
Pear-shaped	62	19.74
Elliptic	22	7.01
Total	314	100.00

One must exercise caution in determining the class limits so that there is no ambiguity as to which class an observation belongs. In constructing the non-overlapping classes, the following important considerations should be borne in mind:

1. The class interval should be of uniform width and of such size that the characteristic features of the distribution are displayed.
2. The class interval must not be too large that can result in errors in assuming that the mid-point of the interval as is the average value of the class.
3. It must also not be so small to give many classes with zero or very small frequencies.
4. The range of the classes should cover the entire range of the data and the classes must be continuous.
5. As a general rule, the average number of classes should be about 15 and never more than 30 nor less than 6.

As an illustration, consider the frequency distribution of the stem length of 11 leaf scars from a coconut plantation. As indicated above, after recording the data, the first task will be to classify these data with the object of reducing them to a form in which they can be conveniently handled. Table 3.2 presents data on stem length of 11 leaf scars to the nearest one tenth of a meter, i.e. decimetre (dm), recorded from 50 palms. Table 3.3 shows its frequency distribution using six classes.

In this example the shortest stem length measures 3.8 dm and the largest measures 13.4 dm. We therefore, have a range of 3.8 to 13.4 dm, which can be divided into six classes by taking class interval of 1.7 dm. The mid point of each class is taken as the *class value*. The number of observations falling within the limits of a particular class is then the frequency of that class.

Table 3.2. Stem length (dm) of coconut population

Plant No.	Stem length	Plant No.	Stem length	Plant No.	Stem length
1	8.6	18	10.4	35	4.3
2	8.3	19	5.9	36	12.8
3	9.6	20	9.9	37	7.7
4	9.1	21	8.2	38	5.8
5	10.4	22	6.1	39	13.2
6	6.3	23	9.3	40	4.8
7	11.9	24	12.7	41	12.6
8	9.3	25	13.4	42	8.9
9	9.7	26	11.0	43	11.6
10	4.5	27	7.9	44	10.8
11	8.6	28	6.8	45	6.3
12	7.8	29	7.7	46	12.2
13	7.9	30	4.1	47	8.9
14	12.6	31	8.5	48	10.2
15	8.4	32	10.4	49	9.8
16	10.4	33	9.8	50	11.9
17	11.5	34	3.8		

Caution must be exercised in classifying those observations whose values fall on the limits of each class range. For instance, the first class includes observations of 3.5 dm up to 5.1 dm; the value 5.2 dm falls into the second class and similarly 6.9 dm falls into the third class and so on. In general, it will be seen that the distribution is marked by low frequencies in the extreme classes. The frequency increases gradually as one approaches the middle of the distribution, giving the distribution a symmetrical appearance.

The last column in Table 3.3 gives the cumulative relative frequency (RCF), which is the percentage of observations below the upper limit of a given class interval. For example, 66% coconut palms have stem length equal or less than 10.2 dm.

The information contained in the frequency distribution can also be expressed in a graph. This method permits a ready grasp of certain important features such as the most frequent class, trends, which are common to some types of frequency distributions and are discussed below:

Table 3.3. Frequency distribution of stem length in coconut population

Class interval	Class value	Frequency of palms in each class	Cumulative frequency	Relative frequency	Cumulative relative frequency
3.5 – 5.1	4.3	5	5	10.00	10.00
5.2 – 6.8	6.0	6	11	12.00	22.00
6.9 – 8.5	7.7	8	19	16.00	38.00
8.6 – 10.2	9.4	14	33	28.00	66.00
10.3 – 11.9	11.1	10	43	20.00	86.00
12.0 – 13.6	12.8	7	50	14.00	100.00
Total		50		100.00	

Note: *It is not necessary that the first class interval begins with the smallest value and the last class interval ends with the largest value. Some provision for lesser values than the smallest observed and bigger values than the largest value observed could be made, in case of necessity, convenience or for presentation.*

Graphical representation of frequency distribution

Graphical representation is another way of presenting frequency distribution of the data. The histogram and frequency polygon are the two graphical representations of frequency data.

In the histogram, the class interval is along the horizontal axis and to rise, over these intervals (the contiguous ones will be overlapping), columns or rectangles whose heights are proportional to the number or frequency (vertical axis) of individuals falling into each class. The histogram for the frequency distribution data from Table 3.3 is shown in Fig. 3.1A.

The information contained in Table 3.3 can also be expressed in the form of frequency polygon which is obtained by plotting class values as abscissa (horizontal axis) and class frequencies as ordinates (vertical axis) and joining of the points by drawing straight lines. The frequency polygon for data in Table 3.3 is shown in Fig. 3.1B.

From Fig. 3.1A and 3.1B, it is evident that maximum frequencies are at the middle of the range and the class frequencies diminish more or less symmetrically in the direction of the two extremes.

It is conventional that frequency polygon and histogram are drawn only for quantitative data. For qualitative data, the frequency distribution is depicted either in the form of a pie chart or a bar diagram. Using data in Table 3.1, the pie chart and bar diagram are shown in Fig. 3.2A and 3.2B, respectively.

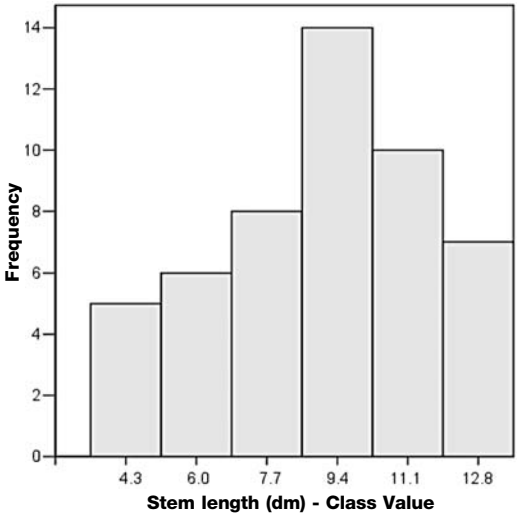


Figure 3.1A. Histogram showing the distribution of stem length in coconut population.

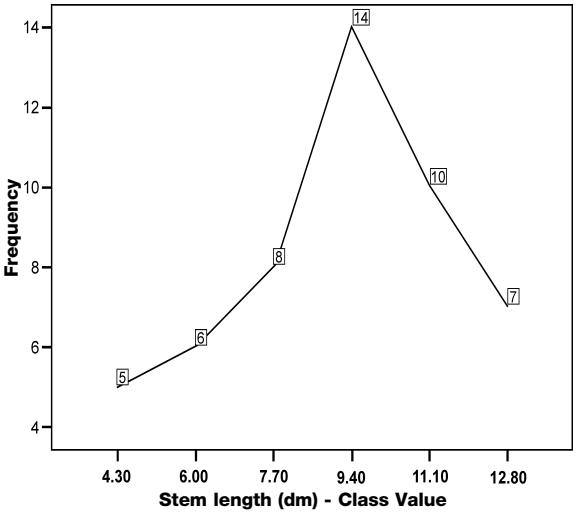


Figure 3.1B. Frequency polygon of stem length in coconut population.

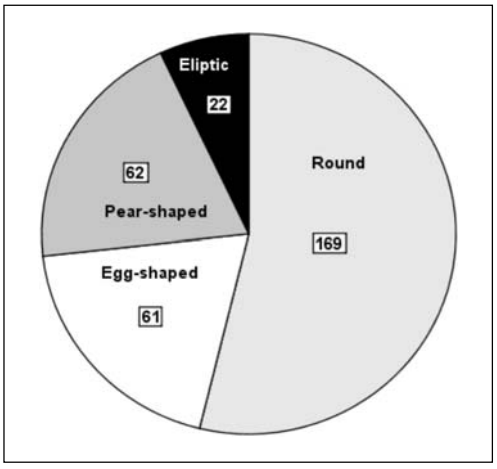


Figure 3.2A. Pie chart for distribution of fruit shape in coconut population.

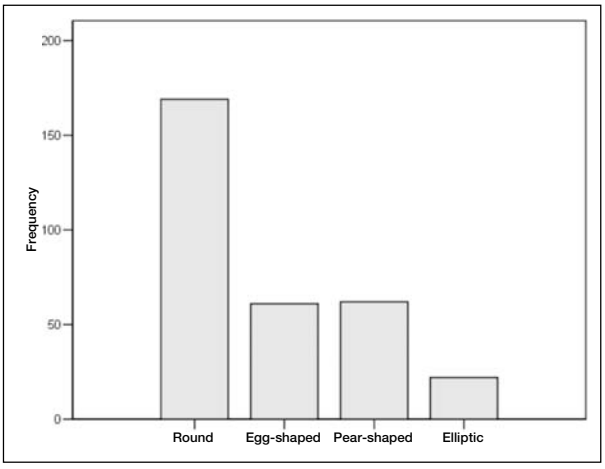


Figure 3.2B. Bar diagram for distribution of fruit shape in coconut population.

Frequency curve and its characteristics

The frequency graph for any data, whether it is *frequency polygon* or a *histogram*, approaches more and more the form of a smooth curve as the number of observations increases and finer class intervals are used. Frequency curves are usually met with a single hump or mode (value with the largest frequency) and can be distinguished from one another by means of the following four characteristics:

1. The central value;
2. The spread of the curve around the central value;
3. The symmetry or the departure from it, termed as *skewness*; and
4. The excess or deficiency of frequencies in the centre and the two extremes compared with the flanks, termed the *kurtosis*.

These characteristics are expressed in terms of the parameters of the distribution *viz*.,

1. Parameters for measures of central tendency;
2. Parameters for measures of dispersion; and
3. Shape parameter for frequency distribution.

The values of the parameters can be computed from the data recorded on all individuals in a population or can be estimated from a sample.

Measures of central tendency

As mentioned earlier, histograms may often appear to have a peak or high region, with the heights of the bars dropping off to zero as one move to the right or left extremes of the histogram. These peaks represent the region of values where a high percentage of the observations fall, and so correspond in some sense to the 'typical' value of the observed character.

This idea of a single 'typical' value, representative of a sample or a population, is very useful in data analysis. Since there are many ways of calculating such a value, each with its own advantages and disadvantages, statisticians use the more generic term 'measures of central tendency'. These measures of central tendency are single numerical values (in most instances) which are intended to indicate the centre or middle region of the distribution of values. Therefore, measures of central tendency are another way of summarizing distribution as typical score, representative score or the point around which the data clusters.

There are three well known measures of the central tendency of a frequency distribution viz., *mean*, *mode* and the *median* as discussed below:

Arithmetic mean

The mean is generally, the arithmetic mean of the values of the individuals in the data and is the most useful measure of central tendency. It is obtained when the sum of the values of the individuals is divided by the total number of individuals. The mean is usually denoted by the symbol μ.

Thus, $\mu = (\Sigma x)/N$, where x denotes the observation, N denotes the number of observations, and Σx denotes the sum of all the observations. For grouped data,

$$\mu = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Where, x_i is the class value of the i^{th} class, f_i denotes the frequency of the i^{th} class and i=1,2,…,k where k denotes the number of classes.

Note: *This formula assumes that all the observations in any class are concentrated at the middle of the class interval. This assumption is of course not strictly true and the formula may therefore give results different from those obtained directly from the individuals without grouping, but the difference is generally negligible.*

Example

For the ungrouped data in Table 3.2, the mean can be calculated as follows:

The mean (μ) = (Σx)/N
= (8.6 + 8.3 + 9.6 + ……+ 9.8 + 11.9)/50
= 452.6/50
= 9.052 dm

For the grouped (frequency distribution) data in Table 3.3, the mean can be estimated as follows:

$$\mu = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

$$= \frac{(4.3 \times 5) + (6.0 \times 6) + (7.7 \times 8) + (9.4 \times 14) + (11.1 \times 10) + (12.8 \times 7)}{(5 + 6 + 8 + 14 + 10 + 7)}$$

$$= \frac{451.3}{50} = 9.026 \text{ dm}$$

You may observe that the effect of grouping on the estimation of the mean is negligible.

Note: *The value of arithmetic mean can be influenced by a relative few highly unusual observations and in some situations; this can give a misleading result.*

Weighted mean

The computation of weighted arithmetic mean is similar to that described for frequency data. If the weights add up to 1, the sum of products of weights and values will give the weighted average.

Note: *The other types of means, used in certain cases, such as geometric and harmonic means are not discussed in this manual.*

Mode

Similarly, the mode can be computed for grouped data. In a frequency table, the modal class is the class that has the greatest frequency. This class can be determined at once from a glance, but the actual value of the mode will be located somewhere in that class interval, not necessarily at the mid-point of the class.

Example

For the ungrouped data in Table 3.2, the value 10.4 occurs most frequently than any other value; as such its modal value is 10.4 dm.

Note: *In the case of small sample data from a continuous distribution, this type of calculation for mode could result in none of the values or an unacceptable value for the mode due to the reason that the sampling fraction happens to be very small and the sample values are likely to be different from one another.*

For the grouped data in Table 3.3, the mode can be calculated as follows:

The class interval 8.6-10.2, with its mid-value as 9.4, is the modal class as this class has the highest frequency of 14 among these class intervals. As indicated above, the actual value of the mode will be located some where in this class interval not necessarily at the mid-point of the class. The differences between the modal frequency and the frequencies in the next lower and the next higher classes (in this case 14 - 8 = 6, and 14 - 10 = 4) are used for interpolation for the exact value of mode (Mo). The exact modal value thus will be calculated as:

$$Mo = B + \left(\frac{D_L}{D_L + D_H} \right) CI$$

Where, B is the starting value of the modal class, D_L and D_H are the differences between the modal frequency and the frequencies of the next lower and the next higher classes, respectively, and CI is the class width. Substituting values,

$$Mo = 8.6 + \frac{6}{6+4} \times 1.7 = 9.6 dm$$

Median

The median is the value, which is located in the middle of a series when the observations are arranged in order of increasing/decreasing magnitude. It divides the series into two halves, half the number of the observations lying above it and half below. The determination of the median is a simple matter when there is odd number of observations in the series. Thus, if 101 observations are placed in the order of their magnitude, the 51st observation will be the value of the median. If there is an even number of observations in a series, the average of the two central values may be taken as the median.

Example

The data in Table 3.2 when arranged in ascending order will be as: 3.8, 4.1, 4.3, 4.5, 12.7, 12.8, 13.2 and 13.4. Since the number of observations in this example is even, the mid-value will be the average of two central values, i.e. observation numbers 25 and 26 in the ascending/descending order. In this example the central values are 9.1 and 9.3, therefore, the median will be (9.1 + 9.3)/2 = 9.2 dm. Suppose in the above example, the last observation, i.e.13.4, was not there, leaving a total number of 49 observations, then the 25th observation, i.e. 9.1, will be the median.

For the grouped data, the median lies in the median class. This class includes the middle value determined as the middle of the array of all observations or the observation at the N/2 position. Hence for the data in Table 3.3, the median lies in the class interval with class value 9.4, as there are 19 observations lower and 17 observations higher than this class. Since N=50, the mean of the 25th and 26th values is the median. The class which includes this value is the median class. Noting that the number of observations with values lower than the median class are 19 (F_1) bringing the total number of observations up to and including the median class as 33 (F_2), and that the class interval is 1.7 dm, we find by interpolation the value of the median in the interval 8.6 to 10.2. The median (Md) value for grouped data is obtained as:

$$Md = B + \frac{(N/2) - F_1}{F_2 - F_1} \times CI$$

Where, B is the lower limit of the median class, F_1 is the cumulative frequency of the class before the median class and F_2 is the cumulative frequency of the median class and CI is the class width.

For the data shown in Table 3.3, the median is obtained as:

$$Md = 8.6 + \frac{(25 - 19)}{(33 - 19)} \times 1.7 = 9.329 \text{ dm}$$

Of these measures of central tendency, the arithmetic mean is by far the most important and commonly used. The mode is the most striking measure of central

tendency. Both the mode and the median are not affected by the extreme values, as is the case with arithmetic mean. Nevertheless, the mean is preferred since it provides more information about the central tendency than the other measures. However, it must be noted that, in a set of values, the mode is the most frequently occurring value; the median is the middle value; and the mean is the average value. No single measure of central tendency provides a complete picture of the data. Suppose the data are clustered in three areas viz., half around a single low value, and half around two large values, both *mean* and *median* may return in the relatively empty middle, and *mode* may return the dominant low value. In such situations, all the three measures of central tendency give a wrong picture of the population. For a normally distributed population, the mode, median and mean are equal.

Measures of dispersion

By dispersion we mean, overall to what extent the data values differ from each other. The mean, as discussed above, gives us an idea of the central value around which the individual observations are distributed; but it tells us nothing about how they are distributed. Two populations can have different patterns of their individual observations, though their means are the same.

Consider the two populations each consisting of five observations. Let A={2, 1, 6, 3, 5, 4} and B={5, 4, 5, 5, 6, 5}. Both populations have mean equal to 5. However, the values in A tend to vary more compared to those of B. We say that A is heterogeneous while B is homogeneous. There are several measures available for measuring the spread or dispersion of observations in the population. These include the range, the interquartile range, standard deviation, variance and the coefficient of variation.

Range

The range of a distribution is the difference between the largest value (maximum) and the smallest value (minimum) in the data set. The range, although a quick measure, gives only a rough estimate of the amount of the variability present in the population. It depends entirely on the two extreme values and takes no account of the pattern of variation among the other values.

Example

For the ungrouped data in Table 3.2, which varies from 3.8 to 13.4 dm, the range is equal to 13.4 - 3.8 = 9.6 dm. In most cases the range is simply presented by indicating the smallest and largest observations, e.g. 3.8 to 13.4.

Note: *Grouped data cannot provide a measure of range.*

Mean of the absolute deviation

Another measure of dispersion is the mean of deviation. This is calculated by adding the absolute values of the deviations of individual observations from their arithmetic mean and then dividing the sum by the number of observations

or $$\frac{\sum_{i=1}^{N}|x_i - \mu|}{N}$$

Where, μ is the population mean and N is the number of observations.
For grouped data, the mean deviation is obtained as:

$$Mean\ deviation = \frac{\sum_{i=1}^{k} f_i|x_i - \mu|}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i}$$

where, x_i is the class value of the i^{th} class, f_i is the frequency of the i^{th} class, μ is the population mean and N is the number of observations.

Example

For the grouped data in Table 3.3 the mean value, as calculated above, is 9.026 dm. These calculations are shown in the Table 3.4 below.

Table 3.4. Estimation of mean deviation for grouped data

Class interval	Class value	Frequency (f_i)	Deviation (d_i)	$f_i d_i$
3.5 – 5.1	4.3	5	4.726	23.630
5.2 – 6.8	6.0	6	3.026	18.156
6.9 – 8.5	7.7	8	1.326	10.608
8.6 – 10.2	9.4	14	0.374	5.236
10.3 – 11.9	11.1	10	2.074	20.740
12.0 – 13.6	12.8	7	3.774	26.418
Total		50		104.788

Thus, the mean deviation is 104.788/50 = 2.096 dm.

For the ungrouped data in Table 3.2, the mean deviation works out to be 2.046 dm, since the sum of the deviation of the observations from the mean value without taking the sign into consideration is 102.296, as explained in Table 3.5.
Thus, the mean deviation works out to be:

Mean deviation = 102.296/50
= 2.046 dm

Table 3.5. Estimation of mean deviation for ungrouped data

Stem length	Deviation (d_i)	Stem length	Deviation (d_i)	Stem length	Deviation (d_i)
8.6	0.452	10.4	1.348	4.3	4.752
8.3	0.752	5.9	3.152	12.8	3.748
9.6	0.548	9.9	0.848	7.7	1.352
9.1	0.048	8.2	0.852	5.8	3.252
10.4	1.348	6.1	2.952	13.2	4.148
6.3	2.752	9.3	0.248	4.8	4.252
11.9	2.848	12.7	3.648	12.6	3.548
9.3	0.248	13.4	4.348	8.9	0.152
9.7	0.648	11.0	1.948	11.6	2.548
4.5	4.552	7.9	1.152	10.8	1.748
8.6	0.452	6.8	2.252	6.3	2.752
7.8	1.252	7.7	1.352	12.2	3.148
7.9	1.152	4.1	4.952	8.9	0.152
12.6	3.548	8.5	0.552	10.2	1.148
8.4	0.652	10.4	1.348	9.8	0.748
10.4	1.348	9.8	0.748	11.9	2.848
11.5	2.448	3.8	5.252		
Grand Total (d_i)					**102.296**

Variance and standard deviation

The standard deviation, denoted by σ is calculated as the positive square root of variance (σ^2). The variance is obtained as the mean of the squared deviations (d) of each observation from the arithmetic mean and is obtained as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N}$$

Where, x_is are the observations, μ is the population mean and N is the total number of observations.

With grouped data, the variance is obtained as:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{N}}{N}$$

Where, μ is the population mean, x_i and f_i are the class value and the frequency of the i^{th} class, respectively and i=1,2,…,k, the total number of classes.

Note: *The formula for the ungrouped data is obtained by simply replacing f with 1 in the above expression.*

Example

For the ungrouped data in Table 3.2, the variance can be calculated as follows:

$$\sigma^2 = \frac{(8.6 - 9.052)^2 + (8.3 - 9.052)^2 + ... + (11.9 - 9.052)^2}{50} = 6.373\ dm^2$$

Alternatively, using the formula involving the original values, i.e. x's

$$\sigma^2 = \frac{(8.6^2 + 8.3^2 + ... + 11.9^2) - \frac{(8.6 + 8.3 + ... + 11.9)^2}{50}}{50}$$

$$= \frac{4415.6 - \frac{(452.6)^2}{50}}{50} = 6.373\ dm^2$$

Hence, $\sigma = \sqrt{6.373\ dm^2} = 2.5245\ dm$

For the grouped data in Table 3.3, the variance can be calculated using the computation given in Table 3.6 below:

Table 3.6. Computations for the mean and variance for the grouped data

Class interval	Class value (*x*)	Frequency (*f*)	*f x*	*f x²*
3.5 – 5.1	4.3	5	21.5	92.45
5.2 – 6.8	6.0	6	36.0	216.00
6.9 – 8.5	7.7	8	61.6	474.32
8.6 – 10.2	9.4	14	131.6	1236.04
10.3 – 11.9	11.1	10	111.0	1232.10
12.0 – 13.6	12.8	7	89.6	1146.88
Total		50	451.3	4398.79

σ^2 = (1/50) {4398.79 - $(451.3)^2$/50}
= (1/50) (4398.79 – 4073.4338)
= 6.507 dm^2

Thus,

σ = 2.551 dm

Observations to be noted

- Since the deviations are squared, the variance is always positive.
- Unlike in the case of mean, the difference between estimates of the variance

for grouped and ungrouped data is not always negligible and a correction is required in computing the standard deviation from grouped data with the help of the above formula.

- It should be remembered that the measures of dispersion are expressed in the same units of measurements such as inches, grams, pounds, etc., in which the observations themselves are measured.
- Among these measures of dispersion, σ is the most commonly used.
- The mean deviation is, perhaps, a simpler measure of variability than the standard deviation, but is not easily amenable to algebraic treatment in the way the standard deviation is.
- In fact, most methods of statistical analysis have evolved around the square of the standard deviation or the variance. It has been shown that the variance is the most informative among the measures of dispersion for populations commonly encountered.

Coefficient of variation

The above measures of dispersion are absolute and are expressed in units in which the observations are recorded. A measure of variation, which is independent of the unit of measurement, and is therefore useful for comparison between different populations and for different characters, is provided by the coefficient of variation (CV). The coefficient of variation is expressed as the standard deviation as a percentage of the mean and is obtained as:

$$\mathrm{CV}(\%) = \frac{\sqrt{\sigma^2}}{\mu}(100)$$

The coefficient of variation remains unaltered by a change in scale such as, for example, change of unit of measurement from feet to centimetres, but it is altered by a change of origin, which affects the mean but not the standard deviation. Thus, the **CV** of temperature would be different if the temperature were measured in centigrade from that obtained when the unit of measurement is Fahrenheit. The **CV** of the percentage of plants affected by a disease would not be the same as that of the percentage of plants free from the infection. A population or character with a higher value of **CV** indicates that it is comparatively more variable than the other population. For experiments comparing treatments, the coefficient of variation indicates the precision with which treatments are compared and is an index of the reliability of the experiment. Thus, the greater the **CV** value, the lower is the reliability of the experiment. The **CV** value varies greatly with the type of experiment, crop grown and characters observed. Most of the characters of interest in coconut experiment show coefficient of variation more than 20% implying the requirement of relatively large sample size for coconut experiments.

Example

With regard to data presented in Table 3.2, the mean and standard deviation for the ungrouped data were obtained as 9.052 and 2.5245, respectively. The **CV** is then obtained as follows:

$$CV = 100 \times (2.5245/9.052)$$
$$= 27.89\%$$

As indicated, the change of units/scale will not have any effect on the estimation of **CV.** In the above example, if we measure the length of stem in terms of millimetres in place of decimetres, the mean value will work out to be 905.2 mm and the variance will work out to be 63732.96 mm^2. Thus, the **CV** will be:

$$CV = 100 \times \sqrt{63732.96}/905.02$$
$$= 100\ (252.4539/905.02)$$
$$= 27.89\%$$

However, when we change the origin for measurement, i.e. if we measure the length of stem over and above 5 dm from ground level and work out the **CV**, we will observe that while the mean changes to 4.052 dm, the variance and the standard deviation remain unchanged as 6.373 dm^2 and 2.524 dm, respectively, then:

$$CV = 100 \times 2.524/4.052$$
$$= 100 \times 0.6232$$
$$= 62.32\%$$

As you can see, the two values are not the same.

With regard to data presented in Table 3.3, the mean and standard deviation for the grouped data are obtained as 9.026 and 2.251, respectively. The CV is then obtained as follows:

$$CV = 100 \times (2.551/9.026)$$
$$= 28.26\%$$

Shape parameter for frequency distribution

After obtaining the mean and variance of a frequency distribution, our interest focuses on which side of the mean lies more number of observations. This is indicated by the value of skewness. The pattern in which the frequency decreases on either side of the mean depends upon the value of kurtosis. Knowledge of skewness and kurtosis will help to visualize the shape of a frequency distribution as discussed below:

Skewness

Skewness measures the symmetry of the frequency distribution. This measure is centred at zero. A large positive value for skewness indicates a long right tail (Fig. 3.3A). That is, the low values of the observations are bunched close to the mean but high values extend far above the mean. Extreme negative value for skewness indicates a long left tail (Fig. 3.3B). Obviously for a symmetric distribution, the value of skewness will be near to zero.

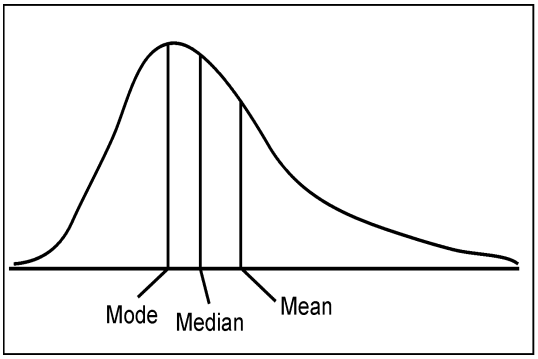


Figure 3.3A. Skewness with positive values.

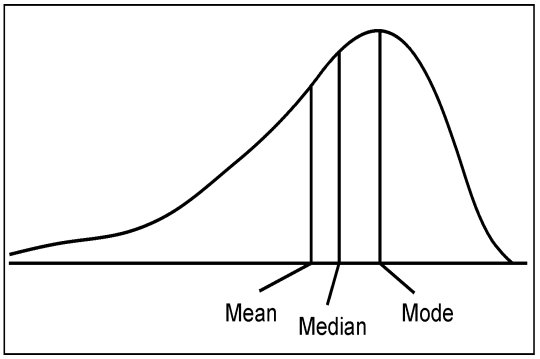


Figure 3.3B. Skewness with negative values.

A measure of the amount of skewness in a population is given by the average value of d^3, where d is the deviation of the observation (or the mid-value of the class, in the case of classified data) from the mean. This quantity is called the *third moment about the mean.* To render this measure independent of the scale on which the data are recorded, it is divided by σ^3, where σ is the standard deviation. The resulting coefficient of skewness is denoted by $\sqrt{\beta_1}$ and sometimes by γ_1.

Thus:

$\sqrt{\beta_1} = \gamma_1 = (\Sigma d^3)/(\Sigma d^2)^{3/2}$ for ungrouped data, and

$= (\Sigma f.d^3)/(\Sigma f.d^2)^{3/2}$ for grouped data.

Example

The computations for skewness using the formula given above can be simplified using the x values, as follows:

First, we obtain the values h_1, h_2 and h_3 which are:

$h_1 = (1/N)\ \Sigma fx$

$h_2 = (1/N)\ \Sigma fx^2$, and

$h_3 = (1/N)\ \Sigma fx^3$

Thereafter, the values m_1, m_2 and m_3 are obtained as:

$m_1 = h_1$

$m_2 = h_2 - h_1^2$, and

$m_3 = h_3 - 3h_1h_2 + 2h_1^3$

Then $\sqrt{\beta_1}$ is obtained as:

$\sqrt{\beta_1} = m_3/m_2^{3/2}$

In the case of grouped data, similar to the calculation of variance as shown above in Table 3.6, a column for fx^3 may be added (Table 3.7).

Table 3.7. Computations for skewness for grouped data

Class interval	Class value (*x*)	Frequency (*f*)	*fx*	fx^2	fx^3
3.5 – 5.1	4.3	5	21.5	92.45	397.54
5.2 – 6.8	6.0	6	36.0	216.00	1296.00
6.9 – 8.5	7.7	8	61.6	474.32	3652.26
8.6 – 10.2	9.4	14	131.6	1236.04	11628.18
10.3 – 11.9	11.1	10	111.0	1232.10	13676.31
12.0 – 13.6	12.8	7	89.6	1146.88	14680.06
Total		50	451.3	4398.79	45330.35

Finding the values

$h_1 = (1/N)\ \Sigma fx = 451.3/50 = 9.026$

$h_2 = (1/N)\ \Sigma fx^2 = 4398.79/50 = 87.9758$

$h_3 = (1/N)\ \Sigma fx^3 = 45330.35/50 = 906.661$

$m_1 = h_1 = 9.026$

$m_2 = h_2 - h_1^2 = 87.9758 - (9.026)^2 = 6.5071$

$m_3 = h_3 - 3h_1h_2 + 2h_1^3 = 906.661 - 3(9.026)(87.976) + 2(9.026)^3 = -4.8755$

Thus,

$\sqrt{\beta_1} = m_3/m_2^{3/2} = -0.2937$

For the ungrouped data in Table 3.2, the skewness can be obtained as:

$\sqrt{\beta_1} = -0.2764$.

Kurtosis

Kurtosis measures the peakness of the distribution. The centre or peak of a distribution is compared against that of a 'normal distribution' for which the kurtosis is zero. Therefore, using the measure of kurtosis, we can state whether the peak of the distribution is much shorter or taller than that of a normal distribution. A large positive value for kurtosis indicates the tails of the distribution are longer than that of normal distribution while a negative value of kurtosis indicates shorter tails (Fig. 3.4).

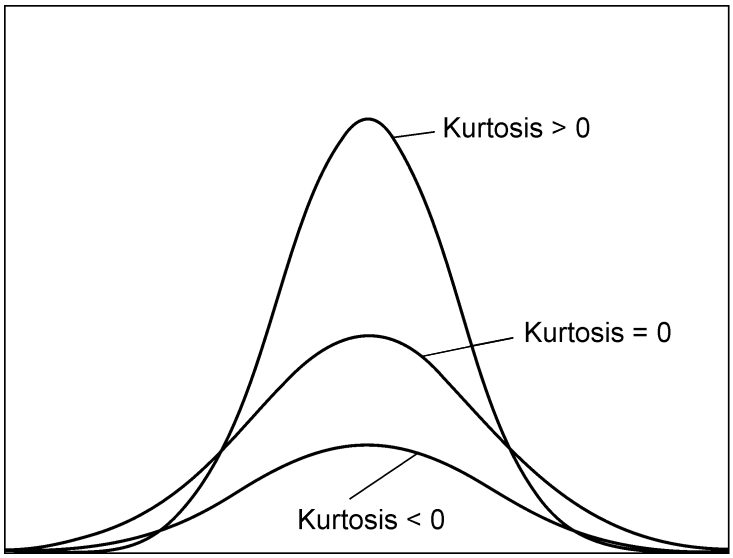


Figure 3.4. Kurtosis with positive (>0) and negative (<0) values.

The coefficient β_2 is used as a measure of kurtosis and is computed by dividing the average value of d^4 by σ^4, where d and σ (defined in the previous section). For the normal distribution, this ratio has the value of 3. As such $\gamma_2 = \beta_2 - 3$ is used as a measure of kurtosis. If the ratio exceeds 3, there is usually an excess of values near the mean and far from it, with a corresponding depletion of the flanks of the distribution curve. Ratios less than 3 result from curves that have a flatter top than the normal.

Thus,

$\gamma_2 = \beta_2 - 3 = \{(\Sigma d^4)/(\Sigma d^2)^2\} - 3$ for ungrouped data, and

$= \{(\Sigma f.d^4)/(\Sigma f.d^2)^2\} - 3$ for grouped data.

Example

Besides the m_1, m_2, and m_3, as defined above, for the calculation of kurtosis we also obtain the values of h_4 and m_4 as indicated below:

$h_4 = (1/N)\sum fx^4$, and

$m_4 = h_4 - 4h_1h_3 + 6h_1^2h_2 - 3h_1^4$

The calculated values for h_4 and m_4 obtained for the grouped data are 9732.491 and 90.609, respectively.

Thus, the calculated value of γ_2 for the grouped data is obtained as –0.860. For the ungrouped data it was -0.655.

Note: *For the calculation of standard deviation, skewness and kurtosis using various statistical packages available, there may be different formulae treating the observations as a sample from a population and making corrections for the sample. Thus, the results are likely to be slightly different for these values, when obtained using different packages. For example, if you are using MS Excel for data analysis for these descriptive statistics, the formulae used are*:

$$\text{Standard Deviation} = \sqrt{\frac{n\sum x^2 - (\sum x)^2}{n(n-1)}}$$

$$\text{Skewness} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

$$\text{Kurtosis} = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Using these above formulae, the values obtained for variance, skewness and kurtosis were: 6.503, -0.285 and -0.595, respectively for the ungrouped data and 6.639, -0.306, and -0.771, respectively for the grouped data in the examples discussed above. Their corresponding values as estimated above are shown in Table 3.9.

Table 3.8. Comparison of estimates using different statistical packages

Estimates	Ungrouped data		Grouped data	
	Calculation	MS Excel	Calculation	MS Excel
Variance	6.373	6.503	6.507	6.639
Skewness	-0.276	-0.285	-0.294	-0.306
Kurtosis	-0.655	-0.595	-0.860	-0.771

The negative value of measure of skewness indicates a left tail for the frequency distribution as can be seen in Fig. 3.3B. However, the value of the skewness in

this example cannot be considered as a very extreme negative value. Similarly, the value of kurtosis is also negative, suggesting that the tails of the distribution is shorter when compared to the normal distribution.

To conclude, the distribution of the data in question is not exactly the normal distribution, but at the same time the measures are not very far from that expected from a normal distribution as well. This is the situation we often come across in practice. The sample we selected for analysis may not be indicating that it is from a normal population and a large deviation once encountered calls for special treatment of the data.

Note: *Because skewness and kurtosis are sensitive to anomalies in distribution, one should study them in conjunction with a histogram. This is because the coefficient of skewness and kurtosis are influenced by the extreme values in the sample while a complete description of the data is made available by means of a histogram.*

Normal distribution

We noted earlier that as the number of observations used for the frequency distribution increases and the class interval is reduced, the graph approaches more and more the form of a smooth curve known as the *frequency curve*. The concept of frequency curve is of great value in statistics, since it provides an excellent summary of the data and reflects their characteristics depending only on a few constants, called parameters, representing the population. Depending on the nature of the variable, whether discrete or continuous, the distributions are termed as discrete or continuous distributions, respectively.

As we discussed earlier, the sample is used to represent the population. The sample parameters hold certain relationships, depending upon the distribution of the population, with the parameters of the population. Mean is by far the most important characteristic and is mostly used in drawing inferences. The distribution of the sample means is of special interest, particularly in biological sciences, where it is seen that the sample mean of characters is distributed nearly as a normal distribution.

Note: *Other continuous distributions, which are extensively used by the biological workers, include student's t-distribution, χ^2 - distribution, F-ratio, specially, for the purposes of testing of hypotheses. Some of the commonly encountered discrete distributions in biological sciences are the Binomial and Poisson distributions. The utility of these distributions are for conducting statistical tests and will be discussed appropriately in subsequent chapters.*

Normal curve

Normal curve or distribution (Fig. 3.5), which is a continuous one, is by far the most important curve in the application of statistical theory to a large variety of biological data. It is determined by two parameters viz., mean (μ) and the standard deviation (σ). It is a bell shaped curve symmetric around the mean. The maximum

ordinate, that is the peak, is at the mean. Thus, the mean coincides with the mode and median in this curve.

As the deviation of a given value on either side from the mean is increased, fewer and fewer values will be lying beyond that of a given value. Nearly 67% of the observations lie within the range $\mu \pm \sigma$ and nearly 95% of the observations lie within the range $\mu \pm 2\sigma$.

For a normal curve, the fraction of the area or observations that get cut off depends only on the ratio $y = (x - \mu)/\sigma$, where x stands for a given observation, σ its standard deviation, and $x - \mu$ its deviation from the mean. This ratio is called the *standard normal deviate*.

It is often our interest to find out the probability of getting values beyond a specific value, for which we use the table of normal probability integral. It may be recalled here that the area under the curve of any probability distribution is unity as the total probability is always equal to 1. The table of normal probability integral gives the fraction corresponding to the area lying to the left of different positive values of the standard normal deviate (and is referred as the ordinates). This area represents the probability of a standard normal deviate being less than the value y ($y > 0$), and consists of the entire area left to the central value zero (i.e. 0.5) and area of the right side leaving the right tail area (probability of the right tail is conventionally denoted as $\alpha/2$). In other words, the probability of values less than y is $0.5 + 0.5 - \alpha/2 = 1 - \alpha/2$.

For negative values of $(x - \mu)/\sigma$, the area to the left of the ordinate at y is given by $\alpha/2$ the 'right tail' area with respect to the corresponding positive value of $(x - \mu)/\sigma$. The area or the frequency of observations lying outside the $\pm$ value of the normal deviate is α and are tabulated in many reference books on statistics (e.g. Fisher and Yates 1963).

Example

As an example for a normal population of the stem length at 11 leaf scars of coconut palms with mean value 9.05 dm and variance 6.25 dm^2 or standard deviation 2.5 dm, the probability of occurrence of stem lengths of 13.85 dm and above can be worked out as follows:

$$y = (x - \mu)/\sigma$$
$$= (13.85 - 9.05)/2.5$$
$$= 1.92$$

In the table of normal probability integral values corresponding to 1.9 and 2.0, values provided are 0.97128 and 0.97725, respectively. Hence, we interpolate the value corresponding to 1.92 as $0.97128 + 2 * (0.97725 - 0.97128)/10 = 0.97247$.

Therefore $1 - \alpha/2 = 0.97247$

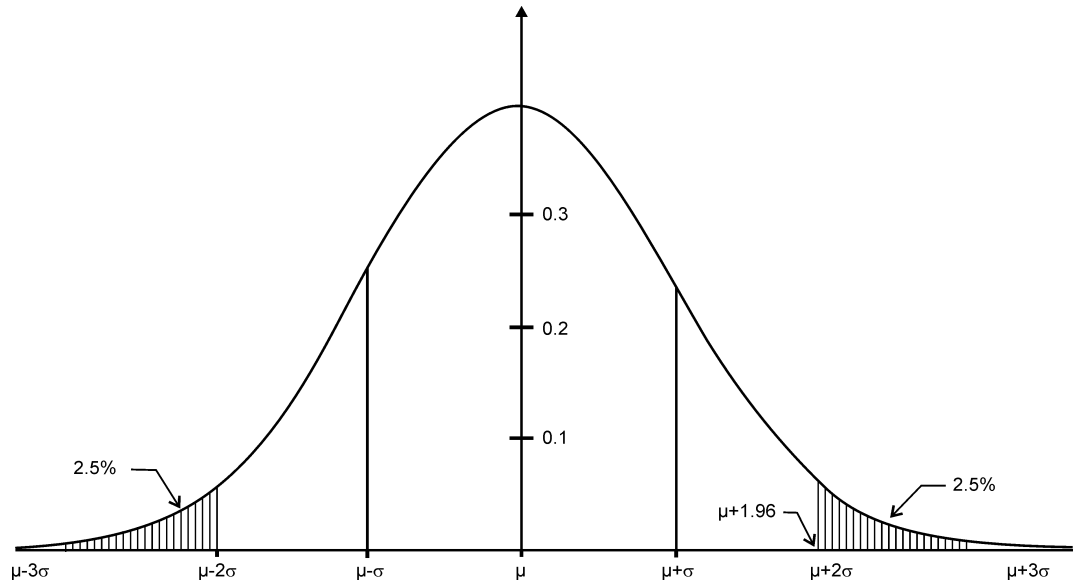


Figure 3.5. Normal distribution with mean m and standard deviation σ.

Thus, the probability of values less than 13.85 dm is 0.97247 and the probability of occurrence of plants with stem lengths of 13.85 dm and above is 1 - 0.97247 = 0.02753 or the proportion of plants with stem length of 13.85 dm and above is 2.75%.

Alternatively, the table for values for normal deviate (y) for chosen level of P (or α in our nomenclature) may be used (Fisher and Yates 1963). From this table, we observe that the P values are provided against y = 1.96 and y = 1.8808 as 0.05 and 0.06, respectively. Thus, the value of P (or α) for y = 1.92 works out to be = 0.05505.

Hence $\alpha/2 = 0.05505/2 = 0.02753$

Therefore, $1 - \alpha/2 = 1 - 0.02753 = 0.97247$ as obtained by using the table of standard normal deviate.

Estimators of mean and variance

The normal distribution depends only on two parameters viz., mean (μ) and standard deviation (σ). The best estimators of the parameters μ and σ^2 based on a sample of n observations $x_1, x_2, \ldots, x_n$ are:

Estimator of μ = Sample mean = $\bar{x} = (1/n)(x_1 + x_2 + \ldots + x_n)$, and

Estimator of $\sigma^2 = s^2 = \{\sum(x_i - \bar{x})^2\}/(n-1)$

These estimates are considered very satisfactory when dealing with normal and nearly normal populations.

Confidence interval (CI_α)

In biological sciences, the most important parameter of interest is the mean value of the character under study. The confidence interval for the population mean is the range of an interval around which the average ($\bar{x}$) of the sample include the 'true average' (μ) of the population with probability 1 - α, [generally, α = 0.05 (i.e. 5%)]. It is taken as $2b_\alpha$ and the *confidence limits* as ($\bar{x} - b_\alpha$, $\bar{x} + b_\alpha$). The value 1 - α is known as *confidence coefficient*. The confidence interval and limits for other population parameters can be similarly defined.

Example

Consider the data given in Table 3.2 as a sample drawn from a normal population with unknown mean and known variance of 6.25. The sample mean based on n = 50 observations was 9.052.

From the normal probability integral we find that the area left of the ordinate for $y = (x - \mu)/\sigma = 1.96$, is 0.975. That is the probability of $\sqrt{n}\ (\bar{x} - \mu)/\sigma \leq 1.96$ is 0.95. In addition, we know that the sample mean is distributed as normal with mean μ and standard deviation $\sigma/\sqrt{n}$ = 0.3535 dm.

Substituting the values of $\bar{x}$ and σ, the confidence limits of the population mean μ are given by $\bar{x} \pm 1.96\ \sigma/\sqrt{n}$ i.e. 9.052 ± (1.96) (0.3535), i.e. 8.359 and 9.745 with 95% confidence coefficient.

Minimum sample size for estimating the mean

The following three factors have to be considered when choosing a sample size, to measure a character:

1. The desired confidence interval (CI)
2. The coefficient of variation (CV)
3. The cost of the sample, which is related to the sample size n

In the normal populations, which are mostly encountered, the value of b_α depends on the standard deviation (σ), sample size (n) and a tabulated factor for the chosen level α.

For example $b_\alpha = 1.96\ (\sigma/\sqrt{n})$, where factor corresponds to α = 0.05. This allows us to calculate the sample size required for a given confidence interval expressed as percentage of mean and confidence coefficient.

Table 3.9 shows the calculated values for the optimal sample size for coconut population according to CV and the desired $(CI_\alpha) = 2\ b_\alpha$. The value of b_α is taken as percentage of μ, the population mean.

Table 3.9. Optimal sample size according to CV (%) and desired $CI_{0.05}$

b_α	Coefficient of Variation (%)								
	5.0	**7.5**	**10.0**	**12.5**	**15.0**	**17.5**	**20.0**	**22.5**	**25.0**
5.0%	4	9	16	25	35	48	62	78	97
7.5%	2	4	7	11	16	21	28	35	43
10.0%	1	3	4	7	9	12	16	20	25
12.5%	1	2	3	4	6	8	10	13	16
15.0%	1	1	2	3	4	6	7	9	11

Example

Suppose the CV of the character stem-length is 24.63%, then a sample size of 97 palms is required to estimate the mean stem length within a difference of 10% of the population mean as seen in Table 3.9 against b_α = 5% and CV = 25%.

Reference

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research (6th edition.). Longman. 146p.

Further reading

Snedecor, G.W. and Cochran, W.G. 1967. Statistical Methods, 6th edition. Oxford and IBH Publishing Co., New Delhi. 593p.

Steel, R.G.D and Torrie, J.H. 1981. Principles and Procedures of Statistics. McGraw-Hill, Singapore. 572p.

Chapter 4: Estimation and tests of significance

A great majority of applied researches involve the comparison of two or more populations. These include experiments designed to discover methods for improved production or performance such as fertilizer trials, variety trials, technology adaptation experiments to compare the mean yields or surveys to determine the phenotypic variability of different populations. To illustrate, consider an experiment to determine which of three treatments is most efficient in increasing the total nut production. The problem of determining whether copra production of different coconut varieties differ from each other is another important application. In general, the interest is to compare the average values of the various characters in different populations assuming the data comes from random samples of units. Sometimes rather than the means, the variability existing in the populations is the parameter of interest. In all these situations, a null hypothesis asserting equality of the parameter to some specific value is tested against an alternative which asserts other values aside from that specified. Other parameters of interest include population proportion, P, variability or measures of association between the characters. For each, an appropriate statistic with known probability distribution is computed. Depending on the value of the statistic, decisions are made according to some criterion. Such a procedure is called test of significance or test of hypothesis. In this chapter, we discuss the most important aspects of statistical inference viz., estimation and tests of significance.

Estimation

The main objective of statistical inference is to estimate the unknown parameters of the distribution of the characteristics of interest and make statements about these parameters. To do this, a random sample is obtained from the population under study and observations are made on each unit. The parameter is then estimated from these observations using the appropriate *estimator* or statistic. The estimator is a rule or formula for obtaining an estimate of the parameter from the sample observations. Estimation is concerned with assessing the magnitudes of the parameters of the populations under study. It is not enough to establish a difference between two treatment means as significant. For operational decisions, it would be necessary to have knowledge of the magnitude of the effect of each treatment. It is also clear that in most cases, we do not have data relating to the population as a whole but only on a random sample of units from it, and the assessed value or estimate is likely to deviate from the parameter it is intended to estimate. This deviation in nature could be due to bias or chance error or both. Therefore, we would like the deviation to be free from bias and the error as small as possible. There are a number of methods for estimating a particular parameter or a function of it, like σ or σ^2. These different methods are called estimators, which are functions of the sample observations.

Estimators for specific parameters are not necessarily unique. Needless to say the quality of the estimates will depend on the quality of the estimators themselves. Hence, not only are we faced with estimating the parameter using a rule but we are also faced with the problem of determining the best estimator. Some desirable properties of estimators include: *unbiasedness*, *consistency* and *efficiency.*

An estimator is said to be unbiased if the mean of the estimates from many samples is equal to the true value of the parameter being measured. This simply means that with repeated estimations of the parameter using independent samples, the average value of the estimates should approach the true value of the parameter. Consistency, on the other hand, refers to the property of an estimator such that as the sample size increases, the estimated value should approach the parametric value more closely and more often. Such an estimator is called *consistent*. The property of *efficiency* is based on the variance of the estimator. An estimator $\hat{\theta}$ based on n observations is said to be efficient if its variance in large samples is least among all the estimators of similar type.

The estimators discussed above are known as *point estimators*. A second type of estimators is known as *interval estimators*. Point estimators give a single value as an estimate of the parameter while interval estimators give a range of values within which the true value may be included. If some probability level which reflects the confidence that the true value is contained in the interval is attached to it, then we call the interval a confidence interval.

The variance of the observations in the sample $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ is an example of a biased point estimator of σ^2 since the expected value of $\hat{S}^2$ is $\sigma^2(n-1)/n$. Hence an unbiased estimator of σ^2 is $\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$. In other words, while calculating the variance of the sample observations, n is used as denominator, whereas, for estimating population variance from the sample, the denominator is (n-1). We refer s^2 as the 'sample variance'.

Similar changes will be required for calculating the population parameters and estimating them from the sample, as we have indicated earlier for the parameters for skewness and kurtosis in chapter 3.

Tests of significance

In statistical inference, we make statements about the parameter of the population of interest. This statement is known as a statistical hypothesis. A statistical hypothesis is a statement about the values of one or more parameters of the population. The hypothesis to be tested is called the null hypothesis (H_0) and the hypothesis to be accepted as true in the event that the null hypothesis is rejected is called the

alternative hypothesis (H_1). The null hypothesis asserts equality or no difference between the populations compared while the alternative hypothesis asserts difference. Hence, in comparing the production performance of two open-pollinated tall populations of coconut, a possible pair of hypotheses is:

H_0: The yield of the two populations is not different from each other versus either one of the following alternative hypothesis
H_1: Population A and Population B yields are different from each other;
H_2: Population A yields less than Population B; or
H_3: Population A yields better than Population B.

An experiment is conducted and based on the value of the test statistic, the hypothesis is rejected or accepted with known probability of error. During the course of testing the hypothesis, there are two situations where the experimenter is likely to commit an error. When the hypothesis is actually true but is rejected, a Type I error is committed. On the other hand, when the hypothesis is actually false and is accepted, a Type II error is committed.

Obviously, one would like to minimize both these errors, which is not possible to achieve simultaneously. Therefore, the magnitude of the Type I error is fixed. For this, as a matter of convenience, the probability level of 5% (0.05) or 1% (0.01) is commonly used (known as *level of significance*) for the cases of rejection of the hypothesis when it is actually true.

The nature of alternative hypothesis (H_1), if the null hypothesis (H_0) was not true, is also to be considered to determine the nature of the test. Keeping the H_0 and H_1 in mind, a test statistic that will help to minimize the second type of error or maximize the power of the test, measured as (1 - probability of Type II error) is developed.

The test of significance is based on the sample data generated/collected. It also depends on the inherent distribution of the population(s) from which the sample data are generated, besides the null and alternative hypotheses. The statistical procedure involves the selection of a suitable test statistic which is a function of the sample observations. Being a function of the observations, the distribution of the test statistic is affected by the distribution from which the observations were sampled. Hence the distribution of the test statistic is determined thereby allowing calculation of probabilities of values of the test statistic. Hence for a given level of significance, one can determine the critical value of the test.

When the result is not significant, we do not say the hypothesis H_0 is accepted. The reason is that, in observational sciences, no finite amount of experimentation or observation can prove or establish a hypothesis. Observations are capable only of disproving or rejecting a hypothesis.

A large difference between the means of two treatments is unlikely to have arisen purely due to chance without a real difference in the effect of treatments. A small difference could more easily have arisen by chance. However, it is necessary to concede that with larger samples the probability of the difference of same small

magnitude arising purely due to chance gets less, and thus one might be in a position to reject the null hypothesis. If, with reasonable amount of data, the null hypothesis is rejected one reaches a definite conclusion.

If the null hypothesis is not rejected by the test, one may conclude either that the null hypothesis is true or that the data do not provide evidence for the null hypothesis to be rejected.

Some of the more important and useful tests for the analysis of coconut experimental data include the t-test for comparing two populations, the F-test for comparing two or more populations and the χ^2- test for testing goodness-of-fit and independence between two variables.

Test of hypothesis about the mean of one population (t-test)

This test, a landmark in the establishment of statistical methods for small samples, was developed by W.S. Gosset, a brewery chemist who wrote under the pseudonym, 'Student' in 1908 and hence has been referred to as "Student t-test". Consider a sample of n observations $x_1, x_2, \ldots, x_n$ drawn from a normal population with mean μ and variance σ^2, i.e., $x \sim N(\mu, \sigma^2)$. The sample mean, $\bar{x}$, is distributed as normal with mean μ, and variance σ^2/n, *i.e.* $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Thus $\sqrt{n}\frac{(\bar{x}-\mu)}{\sigma}$ will be distributed as *N (0,1)*. However, the variance is usually unknown and needs to be estimated from the sample itself. 'Student' worked out the distribution of the test statistic as, $t=\sqrt{n}\frac{(\bar{x}-\mu)}{s^2}$ where, s^2 is the estimated variance from the sample and $t > 0$, as the t – distribution with (n-1) df.

As in the case of the standard normal distribution, the t- tables are also available giving the $P(|t_0| \geq t)$ for a given sample size and degrees of freedom (Fisher and Yates 1963). The t- values required for significance at the 5 and 1 percent levels of significance, for (n-1) df are denoted as $t_{0.05, (n-1)}$ and $t_{0.01, (n-1)}$, respectively. Thus, $t_{0.05, (n-1)}$ is that value such that $P(|t_0| \geq t_{0.05, n-1}) = 0.05$.

It may be noted that the t-distribution is defined only for positive values of t where as the difference ($\bar{x} - \mu$) can take both positive and negative values. Thus, $t_{0.05, (n-1)}$ is that value which is exceeded with a probability 0.025 in the negative direction and with a probability 0.025 on the positive direction making the total probability of 0.05. A test of this kind is therefore referred as two-tailed test. It may be noted here that in Table of t-values, the level of significance refers to the two-tailed test (if not mentioned otherwise). For example, if the level of significance is 5%, the values in the table referred under 5% is with respect to a two-tailed test. If one-tailed test at 5% level is desired, one has to refer under 1% in such Table. It may be noted here that in certain statistical software we may have to specify the complement of the level of significance (i.e. 1-α) with regard to the

'right tail' of the distribution only. For two-tailed test the probability of the right tail is $\alpha/2$. Therefore, if the two-tail test is at 5%, one has to specify 1 - 0.05/ 2 = 0.975 and for corresponding one-tail test, it is 1 - 0.05 = 0.95.

As the df increases, the t-values approaches the values of the standard normal distribution. Thus for large size samples ($n \geq 30$), the tests are carried out taking the distribution of the ratio above as approximately normal and using the normal probability integral tables.

Application: Testing the hypothesis that a population mean is equal to some specified value μ_o.

Data: A sample of n observations ($x_1, x_2, \ldots, x_n$).

Assumptions: The sample is drawn from a normal population.

Hypothesis: Test $H_0 : \mu = \mu_o$ against $H_1 : \mu \neq \mu_o$

Computation:

$$t_c = \sqrt{n}\frac{(\overline{x} - \mu)}{s} \sim t_{\alpha,(n-1)}$$

where, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean,

and $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$ is the sample variance.

Decision: Reject H_0 if $|t_c| > t_{\alpha,(n-1)}$ which is obtained from the t-table. Critical value for 'two-tailed test' at chosen level of significance (α) will be provided in the t-Table. if $|t_c| \leq t_{\alpha,(n-1)}$ the test fails to reject H_0.

Example

It is known that the average copra yield per nut in WCT palms is 172 g. However, a farmer claimed to have WCT palms with average copra yield per nut above 172 g. The plant breeding group wants to test the farmer's claim on superior WCT palms ($\alpha = 0.05$). Samples of 26 nuts were collected from the palms and the weight of copra obtained. The data is given below:

Data: 177.25, 154.50, 173.25, 193.50, 227.50, 155.25, 168.00, 233.00, 150.00, 158.75, 230.00, 200.75, 169.75, 176.75, 158.00, 164.25, 154.50, 162.50, 186.50, 207.00, 250.00, 157.50, 228.50, 216.50, 227.50, 181.50

Hypothesis: $H_0 : \mu = 172$ against $H_1 : \mu \neq 172$

Computation:

$$t_c = \sqrt{26}\,\frac{(\bar{x} - 172)}{s} \sim t_{0.05,25}$$

Compute $\bar{x}$ and s^2 as follows:

$$\bar{x} = \frac{(177.25 + 154.50 + \ldots + 181.50)}{26} = \frac{4862.50}{26} = 187.0192$$

$$s^2 = \frac{\left(177.25^2 + 154.50^2 + \ldots + 181.50^2\right) - \dfrac{(177.25 + 154.50 + \ldots + 181.50)^2}{26}}{25} = 945.3296$$

Therefore,

$$t_c = \sqrt{26}\,\frac{(187.0192 - 172.00)}{\sqrt{945.3296}} = 2.491$$

Decision: The chosen level of significance is 5% and df = n-1 = 25. The tabulated t-value is then $t_{0.05,\,(25)}$ = 2.06. Since the calculated t value (2.491) is greater than the tabulated t value, i.e., $t_c > 2.06$, reject the null hypothesis $H_0 : \mu = 172$.

Conclusion: There is some evidence to support the farmer's claim that the average copra weight is greater than 172.

Note: *The above test is two-tailed. More appropriately we can test $H_0 : \mu = 172$ against $H_1 : \mu > 172$ as we are interested to a population that gives more than the average copra yield per nut. The test statistic remains the same, but we take into account the direction of the deviation also in this case. If the t-statistic is negative, we accept the null hypothesis immediately in this case. If positive, for a chosen level of significance, we need to obtain the value which is exceeded with a probability equal to double the level of significance (i.e. with respect to 5% level of significance, we have to refer tabulated values corresponding to 10%). The test that we adopt in such a case is called a one or single tailed test. (The test with regard to $H_1 : \mu \neq 172$ is referred as two-tailed test).*

If the null hypothesis is rejected based on the two-tailed test, the same decision holds good with regard to single tailed test. With regard to the above example the calculated t-statistic (2.491) is compared with tabulated value of t for 1% and 25 df, which is $t_{0.1,(25)}$ = 1.708. Again $t_c > t_{0.1,(25)}$ and we reject the null hypothesis.

Test of hypothesis about the mean difference of two populations

The tests for comparing the mean difference between two populations can be

conducted using independent samples or using paired samples. First we consider the case of two independent samples.

Independent samples

Let μ_1 and μ_2 be the means of population 1 and population 2, respectively. To test hypothesis about the mean difference between two populations, random samples from each population are taken independently. Let x_1, x_2, …, x_{n1} denote the observations on the first sample and let y_1, y_2, …, y_{n2} denote the observations on the second sample. The difference between the two populations is estimated as the difference between the two sample means.

Assumptions:

(1) The samples are drawn from normal population, and
(2) The variance is same for both the populations.

Hypothesis: $H_0 : \mu_1 = \mu_2$, against $H_1 : \mu_1 \neq \mu_2$

Computation:

$$t_c = \frac{\bar{x} - \bar{y}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where, $\bar{x}$ and $\bar{y}$ are the means of the two samples and s_p^2 is the pooled variance obtained as:

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

Where, s_1^2 and s_2^2 are the sample variances and n_1 and n_2 are the sample sizes.

Decision: Reject H_0 if $|t_c| \leq t_{\alpha, n_1 + n_2 - 2}$. Otherwise, fail to reject H_0.

Example

During a field visit, scientists were told that intercropping in coconut garden favors nut production. To verify this, two gardens, one garden which practiced intercropping and another which practiced monocropping were selected. The mean yields of the two gardens were estimated from the average number of nuts per bunch per palm based on the three oldest bunches at the time of observation. Sixteen and eighteen palms were randomly selected from the monocrop and intercropped gardens, respectively. The data are given in Table 4.1.

Table 4.1. Average number of nuts per bunch from two coconut populations

Sample No.	Monocropping garden (x)	Intercropping garden (y)
1	16.3	21.4
2	15.5	13.2
3	27.3	26.8
4	22.6	29.3
5	12.2	17.4
6	18.7	16.3
7	7.3	12.1
8	9.7	9.0
9	21.3	20.8
10	15.5	17.7
11	22.2	19.4
12	13.2	15.2
13	19.0	18.3
14	17.4	18.0
15	28.8	25.4
16	14.9	17.3
17		18.8
18		19.5
Total	281.8	335.9
Sample size	16.0	18.0
Sample Mean	17.6	18.7
Sample Variance	34.6	25.1

Hypothesis: Test $H_o : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Computation:
Sample size: $n_1 = 16$ and $n_2 = 18$

$$s_p^2 = \frac{(16-1)34.56 + (18-1)25.11}{16+18-2} = 29.54$$

$$t_c = \frac{17.62 - 18.66}{\sqrt{29.54\left(\frac{1}{16} + \frac{1}{18}\right)}} = -0.56$$

The df = $n_1 + n_2 - 2 = 32$

Decision: From the table of t-values, we get $t_{0.05, 32} = 2.042$. Since as $|t_c| < 2.042$, we do not reject the null hypothesis $H_0 : \mu_1 = \mu_2$

Conclusion: The data do not provide evidence that intercropping favours coconut production.

With respect to the one-tailed test (against the alternative hypothesis $H_1 : \mu_1 < \mu_2$) also the null hypothesis is not rejected as $t_c < t_{0.1,\ 32} = 1.31$

Paired Samples

Paired samples arise either by the use of similar units or through self-pairing. In the first case, the samples are matched and each member of the pair belongs to one of the populations to be compared. Pairs of coconut fruits where each pair comes from the same bunch can be used to determine the efficacy of two virgin oil extraction methods. On the other hand, self-pairing arises in "before and after" treatments, where the individuals are observed twice, once before the treatment and once after the treatment.

For example, to test if the copra yield in two seasons of harvest of a plantation are equal, a researcher can observe the copra yields of the sampled palms for both seasons or could choose different sets of sampled palms in different seasons.

When paired samples are used, the test of hypothesis about the mean difference, $\mu_D = \mu_1 - \mu_2$, between the two populations is conducted as follows:

Data: Paired observations from a random sample of size n. Observations are denoted by $(x1_1, x2_1)$, $(x1_2, x2_2)$, …, $(x1_n, x2_n)$.

Hypothesis: $H_0 : \mu_D = 0$ (i.e., the averaged paired difference is zero), against $H_1 : \mu_D \neq 0$

Computation:

$$t_c = \sqrt{n}\frac{\bar{d}}{s}$$

Where, $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$ and is the mean of n paired differences $d_i = x_{2i} - x_{1i}$

$$s^2 = \frac{\sum_{i=1}^{n} d_i^2 - \dfrac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}}{n-1}$$

Decision: Reject H_0 if $|t_c| > t_{\alpha,n-2}$, otherwise fail to reject H_0.

Example

To study the effect of certain hormones, 16 palms were selected at random and data on the average number of nuts per bunch before and after the hormone

application were recorded. The paired observations (pre- and post-hormone application) are given in Table 4.2.

Hypothesis: $H_0 : \mu_D = 0$, against $H_1 : \mu_D \neq 0$

Where, μ_D = difference between the means before and after the hormone treatment

Table 4.2. Paired observations for pre- and post- hormone application in coconut

Palm Number	1	2	3	4	5	6	7	8
Pre-treatment	16.3	15.5	27.3	22.6	12.2	18.7	7.25	9.7
Post-treatment	21.4	13.2	26.8	29.3	17.4	16.3	12.1	9.0
Difference (d_i)	-5.1	2.3	0.5	-6.7	-5.2	2.4	-4.85	0.7
Palm Number	9	10	11	12	13	14	15	16
Pre-treatment	21.3	15.5	22.2	13.2	19.0	17.4	28.8	14.9
Post-treatment	20.8	17.7	19.4	15.2	18.3	18.0	25.4	17.3
Difference (d_i)	0.5	-2.2	2.8	-2.0	0.7	-0.6	3.4	-2.4

$$\bar{d} = \frac{\left[(-5.1)+2.3+\ldots+(-2.4)\right]}{16} = -0.9844$$

$$s^2 = \frac{\left[(-5.1)^2+2.3^2+\ldots+(-2.4)^2\right] - \frac{\left[(-5.1)+2.3+\ldots+(-2.4)\right]^2}{16}}{15} = 10.1899$$

$$t_c = \sqrt{16}\,\frac{-0.9844}{3.1922} = -1.2335$$

Decision: Since $|t_c| < t_{0.05,15} = 2.131$, we fail to reject H_0.

Conclusion: The data do not give evidence that the number of nuts per bunch is affected by hormonal treatments.

Note: *It has been shown that even in cases where the normality assumption may not be satisfied, the results of t-test hold since the* t-test *is robust to moderate departures from normality. However, the assumption that the paired differences are mutually independent needs to be satisfied. This is satisfied if the paired observations are randomly sampled.*

Test of hypothesis about equality of means of more than two populations

With more than two populations under study, we might wish to test the hypothesis that the samples come from populations having the same mean (i.e. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$) against the hypothesis that the population means vary significantly from

each other. For these problems, the more appropriate procedure is known as analysis of variance (ANOVA). The ANOVA is an extension of the independent samples t-test of two populations. The ANOVA and F-test are extensively used in the analysis of experimental data as such examples are deferred for later chapters. F-test for equality of two variances is described below.

F- test for equality of two variances

The distribution of the ratio of two independent estimates of variance (of a normal distribution) is known as F-distribution. It has two parameters v_1 and v_2 as the DF associated with the estimate of variance in the numerator and denominator. To perform statistical tests, we look for the tabulated F value at appropriate level of significance and for the appropriate parameters, i.e. v_1 degrees of freedom (for numerator) and v_2 degrees of freedom (for denominator).

An application of the F-test is in testing the equality of two variances. Consider two independent random samples of size n_1 and n_2. Let $x_1, x_2, \ldots, x_{n1}$ and $y_1, y_2, \ldots, y_{n2}$ denote the observations on sample 1 and sample 2 , respectively. Assume that the two populations are normally distributed.

Hypothesis: $H_o : \sigma_1 = \sigma_2$, against, $H_1 : \sigma_1 \neq \sigma_2$

Where, σ_1^2 and σ_2^2 are the population variances of the two populations.

Computation:

$$F_c = \frac{s_x^2}{s_y^2}, \text{ if } s_x^2 > s_y^2 \text{ or}$$

$$F_c = \frac{s_y^2}{s_x^2}, \text{ if } s_y^2 > s_x^2$$

Where, s_y^2 and s_x^2 are the sample variances

Decision: Reject H_0 if $F_c > F_{\alpha,(n_1-1,n_2-1)}$. Otherwise, fail to reject H_0.

Example

Consider data used in the previous example of comparing monocropping and intercropping (Table 4.1). While testing the equality of means using the t-test, we assumed that the variances of the two populations are the same without really testing for equality of variance. The F-test allows us to validate the assumption of equal variances.

Hypothesis: $H_o : \sigma_1 = \sigma_2$, against $H_1 : \sigma_1 \neq \sigma_2$

Sample variances estimated for monocropping and intercropping gardens are (refer Table 4.1):

S_x^2 = 34.56257 with 15 df

S_y^2 = 25.11075 with 17 df

Since $s_x^2 > s_y^2$,

Computation:

$$F_c = \frac{34.56257}{25.11075} = 1.376$$

Decision: Since $F_c < F_{0.05(15,17)}$ = (2.308), we do fail to reject H_0.

Conclusion: The variances of the two samples did not vary significantly from each other. Hence the assumption of homogeneity of variances is valid.

Test for equality of more than two variances

To test the homogeneity of variances, the test based on chi-square is used. Consider random samples of sizes $n_1, n_2, \ldots, n_k$ drawn independently from normal population. Let the observations of i^{th} sample be denoted as $x_{i1}, x_{i2}, \ldots, x_{inj}$.

Hypothesis: $H_0 : \sigma_1 = \sigma_2 = \ldots. = \sigma_k$, against H_1 : at least one σ_i is different from the rest

Computation:

$$\chi_c^2 = 2.3026 \left\{ \frac{\left(\sum_{i=1}^{k} (n_i - 1) \right) \log_{10} s^2 - \sum_{i=1}^{k} (n_i - 1) \log_{10} s_i^2}{C} \right\}$$

where,

$$s_i^2 = \frac{\sum_{i=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i=1}^{n_i} xij \right)^2}{n_i}}{n_i - 1}$$

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

$$C = 1 + \left\{\frac{1}{3(k-1)}\right\}\left\{\sum_{i=1}^{k}\frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{k}(n_i - 1)}\right\}$$

The coefficient 2.3026 is included in the formula because logarithm to the base 10 is being used. If natural logarithms are used, this coefficient should not be included in the expression.

Decision: Reject H_0 if $\chi_c^2 > \chi_{\alpha,k-1}$. Otherwise, fail to reject H_0.

Example

The weights of randomly selected nuts of four coconut accessions are shown in Table 4.3. Test whether the variance of nut weight is the same in all four accessions.

Table 4.3. Weight of nuts(g) from four coconut populations

Sl. No.	Accession 1 (x_{1i})	Accession 2 (x_{2i})	Accession 3 (x_{3i})	Accession 4 (x_{4i})
1	438	1004	1270	770
2	449	1018	1421	775
3	453	1019	1425	784
4	518	1032	1435	786
5	564	1045	1445	788
6	608	1053	1446	790
7	610	1056	1461	791
8	651	1060	1506	795
9	680	1068	1526	802
10	700	1074	1568	806
11		1087	1610	813
12		1095	1780	824
13		1116		838
14		1141		

Computation: The calculation of the test statistic uses the computations shown in Table 4.4.

Table 4.4. Estimates of variance for nut weight in four accessions

Sample size (n_i)	Accession 1 10	Accession 2 14	Accession 3 12	Accession 4 13
$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$	5671	14868	17893	10362
$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^{\ 2}$	3303039	15809326	26851289	8263756
$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^{\ 2} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2}{\sum_{i=1}^{k} n_i}$	87014.9	19510	171334.9	4444.9
Estimates of variance (s_i^2)	9668.3	1500.8	15575.9	370.4

From the table, $\sum_{i=1}^{k}(n_i - 1) = 9 + 13 + 11 + 12 = 45$

$k = 4$

$$s^2 = \frac{9(9668.322) + 13(1500.769) + 11(15575.9) + 12(370.4103)}{45} = 6273.439$$

$$\sum_{i=1}^{k}\frac{1}{(n_i - 1)} = \frac{1}{9} + \frac{1}{13} + \frac{1}{11} + \frac{1}{12} = 0.3623$$

$$C = 1 + \frac{1}{3(4\text{-}1)}\left(0.3623 - \frac{1}{45}\right) = 1.0378$$

$\sum(n_i - 1)\log s_i^2 = 9(3.9854) + 13(3.1763) + 11(4.1925) + 12(2.5687) = 154.102$

Then,

$\chi_c^2 = 2.3026 \{45 \times \log (6273.439) - 154.102\}/1.037784$
$= 2.3026 \{45 \times 3.797506 - 154.102\}/1.037784$
$= 2.3026 \times 16.78633/1.037784$
$= 37.2438$

with v = k-1 = 3 DF.

Decision: From table of χ^2, we get $\chi^2_{0.05,\ 3} = 7.81$

Since $\chi^2_c > \chi^2_{0.05,\ 3}$, we reject the null hypothesis.

Conclusion: The variance of the nut weight of the four accessions is not homogeneous.

Analysis of qualitative data

Although most of the variables of interest are generally quantitative and continuous in nature, qualitative, categorical or enumerative data also arise in research like characterization studies. Two types of tests will be given here: test of goodness of fit and test of independence.

Test of goodness-of-fit for frequency data

The χ^2 test of *goodness-of-fit* is used to determine whether the observed frequencies agree with the expected or the hypothesized frequencies. The data consist of counts or frequencies of observations falling into *c* classes or categories of a random variable observed on a random sample. The observed frequencies are denoted as $O_1, O_2, \ldots, O_c$ such that $\sum_{1=1}^{c} O_i = n$. These observed frequencies are compared with the expected frequencies $E_1, E_2, \ldots E_c$. If we let *F(x)* denote the hypothesized frequencies and *F(x),* the true frequency distribution of the population, then the hypotheses for the test can be expressed as follows:

Hypothesis: H_0 : F(x) = F(x), against H_1 : F(x) ≠ F(x)

Computation:

$$\chi_c^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{c-1} \text{ under } H_0.$$

Decision: Reject H_0 if $\chi^2_c > \chi^2_{\alpha,\ c-1}$. Otherwise, fail to reject H_0.

Example

The average number of days for germination of seed nuts is known to be 60 days with a standard deviation of 10 days. It was desired to test whether the character follows a normal distribution. Data from 20 randomly selected seed nuts were recorded and ordered according to increasing magnitude as shown below:

46, 47, 48, 48, 48, 49, 52, 52, 54, 54, 55, 57, 65, 65, 66, 67, 69, 72, 73, 76

Hypothesis: H_0 : F(x) = F(x), against H_1 : F(x) ≠ F(x)

Where, F(x) ≈ N (60, 100)

Computation:

Note that $Y = \frac{(X - 60)}{10} \sim N(0,1)$. Hence, 50% of Y values lie between -0.6745 and 0.6745. Also the mean is 0 and 50% of the observations lie below it. Knowing these and $X = 60+10Y$, then under H_0, the following probability statements can be made:

$P[X<(60 + 10(-0.6745))] = F(53.255) = 0.25$

$P[X<(60 + 10(0))] = F(60.0) = 0.50$

$P[X<(60 + 10(0.6745))] = F(66.745) = 0.75$

From the sample we estimate the probabilities. Hence, $F(53.255)$ which is the proportion of observations below 53.255 is 8/20 or 0.40. Similarly, F(60) = 12/20 = 0.6 and F(66.745) = 15/20 = 0.75.

Corresponding to the distribution function as defined under H_0 we then form four classes as shown in Table 4.5.

Table 4.5. Observed frequencies for number of days to germination and expected frequencies when normal distribution is assumed*

Classes	<53.255	53.255–60	60–66.745	>66.745	Total
Observed frequency (O_i)	8	4	3	5	20
Expected frequency (E_i)	0.25 x 20 = 5	5	5	5	20
$O_i - E_i$	3	-1	-2	0	0
$(O_i - E_i)^2$	9	1	4	0	14

*Expected frequency = 20/4 = 5 for all the classes.

Now, $\chi_c^2 = 14/5 = 2.8$ with df c-1 = 3.

Decision: From Table for values χ^2, we get $\chi^2_{0.05, 3} = 7.815$. Since $\chi_c^2 < \chi^2_{5\%, 3}$, we fail to reject the null hypothesis.

Conclusion: The variable 'number of days taken for germination' has normal distribution with mean 60 days and standard deviation of 10 days.

Note: *The use of χ^2 requires that the frequency expected in any class is not too small, i.e. 5 or less. Pooling of frequencies in the adjacent classes can be resorted to satisfy this criterion remembering, however, that the pooled classes need to be treated as one class and this pooling should not be carried out indiscriminately. The test is applicable only in comparing observed and expected values of absolute frequencies and not relative frequencies or proportions. The degrees of freedom is determined after taking into account the number of restrictions like fixed total frequency, number of parameters/constants estimated from the data to obtain the expected frequencies, etc., and reducing the number of final number classes by the number of such restrictions.*

Test for independence

Consider n independent observations of a random variable X classified according to two criteria. Suppose that there are c classes for one factor and r classes for the second factor. Hence the data of n observations can be classified into r x c classes and can be arranged in a table with r rows and c columns. The number of observations in the $(ij)^{th}$ cell (in i^{th} row and j^{th} column) of the r x c contingency table is denoted by O_{ij}. The n observations are drawn at random and each observation is classified into one of categories of the first factor and to one of the categories of the second factor.

H_0: The row factor is independent of the column factor

H_1: The row factor is not independent of the column factor

Computation:

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

Where, E_{ij} is the expected frequency of $(ij)^{th}$ cell under the null hypothesis and obtained as: $E_{ij} = \frac{R_i C_j}{n}$, with R_i the i^{th} row total and C_j the j^{th} column total.

Decision: Reject H_0 if $\chi_c^2 > \chi^2_{\alpha,\ (r-1)\ (c-1)}$. Otherwise fail to reject H_0.

Example

To compare the germination percentage in embryo cultures of 8, 9 and 11 month old coconuts, 179 coconut embryos were tissue cultured. The numbers of germinated and non-germinated embryos for the three different ages of coconuts were recorded and the results are summarized in the two-way table below.

Table 4.6. Germination percentage of embryos at varying ages (months)

	Observed frequencies				Expected frequencies		
	8	9	11	Total	8	9	11
Germinated	28	39	45	112	60x112/179 = 37.542	57x112/179 = 35.665	62x112/179 = 38.793
Not germinated	32	18	17	67	60x67/179 = 22.458	57x67/179 = 21.335	62x67/179 = 23.207
Total	60	57	62	179	60	57	62

One may check that the row total and column total of expected frequencies will be the same as that of observed frequencies.

$$\chi_c^2 = \frac{(28-37.52)^2}{37.52} + \frac{(39-35.665)^2}{35.665} + \ldots + \frac{(17-23.207)^2}{23.207} = 9.965$$

Note that alternatively χ_c^2 can be computed as:

$$\chi_c^2 = \frac{28^2}{37.542} + \ldots + \frac{17^2}{23.207} - 179 = 9.965$$

Decision: We get df =(r - 1) (c - 1) = 2 and $\chi^2_{c\ 0.05,\ 2}$ = 5.991. Since χ_c^2 >5.991, reject H_0.

Conclusion: The percentage germination is not independent of the age of coconuts.

References

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research, 6th edition. Longman, Edinburgh. 146p.

Student. 1908. On the probable error of mean. Biometrika. 6: 1-25.

Chapter 5: Analysis of relationships between variables

In the previous chapter, we have discussed the measurement of variation on a single variable. We shall now consider the simultaneous variation of two or more variables. It often happens that changes in one variable are accompanied by changes in another variable and that a definite relation exists between the two. Correlation and regression analyses are two statistical procedures used for analysis of relationships between variables. We measure the association between two variables by the coefficient of correlation and the functional relationship of one variable with other variable(s) by the regression equation. Other applications of these two methods like path-coefficient analysis provide further information on the direct effects of causative variable on a dependent variable and also its indirect effect through other causative variables. These techniques are discussed in this chapter.

Correlation

In the case of coconut and similar plants, as the crop grows taller, its girth also tends to grow wider. Thus, we say that these two characters viz., the plant height and the stem girth are correlated. When the two variables change together in such a way that an increase in one variable is accompanied by an increase in the other, as shown in the above example, the variables are said to be positively correlated. Should an increase in one variable go hand in hand with a decrease in the other, the variables are said to be negatively correlated. In biological measurements, the relationship is not likely to be so complete in the sense that a certain unit change in the measurement of one variable may not be accompanied by the same degree of change in the other. Thus, the necessity to quantify the relationship arises for which the coefficient of correlation is used.

Pearson's coefficient of correlation

For two variables X and Y with respective means μ_x and μ_y and standard deviations σ_x and σ_y, the coefficient of correlation, usually indicated by the symbol ρ, is defined as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

where, σ_{xy} is the covariance between X and Y and is defined as:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

$$= \frac{\sum_{i=1}^{N} X_i Y_i - \frac{\left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{N}}{N}$$

The correlation coefficient is independent of the units of measurements and its value ranges from -1 to +1. A qualitative description of the magnitude of the correlation coefficients is as follows:

Absolute Value of the Correlation Coefficient	Qualitative Interpretation
0.8 – 1.0	Very strong
0.6 – 0.8	Strong
0.4 – 0.6	Moderate
0.2 – 0.4	Weak
0.0 – 0.2	Very Weak

Estimation of the correlation coefficient

The Pearson's coefficient of correlation (ρ) between two variables X and Y is estimated from a random sample of n observations by

$$r = \frac{S_{xy}}{S_x S_y}$$

where, s_{xy} is the sample covariance of X and Y and s_x and s_y are the sample standard deviations. The sample covariance for ungrouped data is obtained as:

$$s_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{n-1}$$

Whereas, for grouped data it is estimated as:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} f_i - 1}$$

Where, $\bar{x}$ and $\bar{y}$ are the sample means of x and y and f_i is the frequency of the i^{th} class. The standard deviations are estimated using formulas given in earlier chapters. The expression for sample correlation coefficient can be simplified as:

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

Test of hypothesis about the correlation coefficient

Consider observations on two or more variables on *n* units of a randomly selected sample. It is assumed that a linear association exists between the variables.

Hypothesis: $H_o : \rho = 0$, against $H_1 : \rho \neq 0$

Computation:

$$t_c = \frac{r}{\sqrt{\frac{\left(1-r^2\right)}{n-2}}} \sim t_{(n-2)}$$

where, *r* is the sample correlation coefficient. When H_0 is true, the test statistic t_c follows the *t*-distribution with n-2 df.

Decision: Reject H_0 if $t_c \geq t_{0.05(n-2)}$, otherwise fail to reject the null hypothesis.

Note: *This test can be more easily applied with the help of a Table (Table VI, Fisher and Yates 1963), which gives the values of r (irrespective of sign) required for different levels of significance for different df. For selected df, the following Table 5.1 gives values of r at 5% and 1% level of significance.*

Table 5.1. Tabulated values to test the significance of correlation (for selected DF)

DF	1	2	3	5	8	10	15	18	20	25	30	100
5%	0.997	0.950	0.878	0.754	0.632	0.576	0.482	0.444	0.423	0.381	0.349	0.195
1%	1.000	0.990	0.959	0.874	0.765	0.708	0.606	0.561	0.537	0.487	0.449	0.254

Example

Consider data on fruit characteristics of 20 randomly selected West Coast Tall palms shown in Table 5.2. The variables are fruit weight (FW), nut weight (NW), volume of cavity (VC), endosperm/kernel weight (KW) and copra weight (CW). Test the

hypothesis that FW and NW are linearly associated. The various computations required for obtaining the correlation between characters are presented in Table 5.3.

Table 5.2. Fruit characteristics of 20 West Coast Tall (WCT) palms

Palm No.	Fruit weight (FW) (g)	Nut weight (NW) (g)	Volume of cavity (VC) (cm^3)	Kernel weight (KW) (g)	Copra weight (CW) (g)
1	1216	662	180	346	172
2	1445	735	200	383	187
3	786	466	110	262	157
4	784	467	110	272	152
5	750	464	120	262	155
6	1004	638	190	305	194
7	838	505	140	279	170
8	892	560	180	264	165
9	1019	614	190	321	198
10	860	486	170	252	158
11	1060	701	230	362	224
12	928	569	180	305	194
13	1568	875	310	429	245
14	1461	868	300	414	250
15	1141	686	270	386	209
16	1170	722	230	400	206
17	960	548	140	275	162
18	712	437	120	240	144
19	1002	532	130	280	174
20	1183	555	110	286	164

Table 5.3. Computations required for obtaining the correlation between characters

	FW	NW	VC	KW	CW
$\sum_{i=1}^{20} X_i$	20779	12090	3610	6323	3680
$\sum_{i=1}^{20} X_i^2$	22726265	7626284	724300	2066707	694866
$\sum_{i=1}^{20} X_i^2 - \frac{\left(\sum_{i=1}^{20} X_i\right)^2}{20}$	1137923	317879	72695	67690.5	17746
$\sqrt{\sum_{i=1}^{20} X_i^2 - \frac{\left(\sum_{i=1}^{20} X_i\right)^2}{20}}$	1066.7	563.8	269.6	260.2	133.2

Hypothesis: $H_0 : \rho = 0$, against $H_1 : \rho \neq 0$

Computation:
To obtain the coefficient of correlation between FW=X and NW=Y, first multiply the paired values of the variables and sum as:

$$\sum_{i=1}^{n} X_i Y_i = (1216)\ (662) + (1445)\ (735) + \ldots + (1183)\ (555) = 13119918$$

Hence the corrected sum of cross products is:

$$\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 13119918 - \frac{(20779)(12090)}{20} = 559012.5$$

Substituting in the formula for *r*,

$$r_{FW,NW} = \frac{559012.5}{(1066.73)(563.81)} = 0.9295$$

Decision: Since $r_{FW,NW} > 0.444$ (the tabulated value of r with 18 df at 5% level, refer Table 5.1), we reject the null hypothesis. Therefore we say the correlation between FW and NW is significant at 5% level. It can be verified that the correlation between FW and NW is significant at 1% as well.

Alternatively, we may compute the t-statistic for testing the significance of correlation as:

$$t_c = \frac{0.9295}{\sqrt{\frac{(1-0.9295^2)}{20-2}}} = 10.6895$$

Decision: Since $t_c > t_{0.05,\ 18} = 2.101$, we reject H_0.

Conclusion: The fruit weight and nut weight are very strongly positively correlated i.e. an increase in fruit weight is associated with an increase in nut weight.

Correlation matrix

When there are more than two variables, it is more convenient to present the correlation coefficients between all possible pairs of variables in matrix form. To illustrate the coefficients of correlation between all possible pairs of fruit characters

in Table 5.2 were computed and presented in matrix format as shown in Table 5.4. The diagonal entries represent the correlation of the character with itself and thus are all equal to 1. The correlation coefficients are given in the off-diagonal cells. Hence the correlation coefficient between NW and VC is 0.922 while that between CW and KW is 0.894.

Table 5.4. Correlation matrix between fruit characters

	FW	NW	VC	KW	CW
FW	1.000	0.929	0.769	0.888	0.772
NW		1.000	0.922	0.960	0.924
VC			1.000	0.896	0.929
KW				1.000	0.894
CW					1.000

Since all the values in the Table 5.4 are above 0.561 and DF =18, we conclude that all the correlations are significant at 1%.

Testing the equality of two correlation coefficients

To compare two correlation coefficients, the tests of significance involve the *Z* transformation, specified by the relation:

$$Z = (1/2)\ \log_e \{(1+r)/(1-r)\} = (1/2)\ [\log_e (1+r) - \log_e (1-r)]$$

which approaches the normal distribution for all values of the number of pairs *n* with a standard error $\sigma_Z = \frac{1}{\sqrt{n-3}}$

Example

Test whether correlation coefficient between fruit weight (FW) and copra weight (CW) is significantly different from the correlation coefficient between nut weight (NW) and Copra Weight (CW) based on the data provided in Table 5.2.

Based on sample size n = 20, the coefficients of correlation were obtained as,

$r_{FW,CW}$ = 0.772 and $r_{NW,CW}$ = 0.924

Hypothesis: $H_0 : \rho_{FW,CW} = \rho_{NW,CW}$, against $H_1 : \rho_{FW,CW} \neq \rho_{NW,CW}$

Computation:

Obtain the *Z* values as:

$$Z_1 = \frac{[\ln (1+0.772) - \ln (1-0.772)]}{2} = 1.0253$$

$$Z_2 = \frac{[\ln (1+0.924) - \ln (1-0.924)]}{2} = 1.6157$$

Since the sample size is the same, the variance of Z_1 and Z_2 is given by $1/(n\text{-}3)$ which is $1/17 = 0.058824$. Under the null hypothesis, $(Z_1 - Z_2)$ is distributed as normal with mean 0 and variance $v(Z_1) + v(Z_2) = 2\ (0.0588) = 0.1176$.

$$Z_c = \frac{|Z_1 - Z_2|}{\sqrt{v(Z_1 - Z_2)}} = \frac{|1.0253 - 1.6157|}{\sqrt{0.1176}} = 1.7215$$

Decision: Since $Z_c < 1.96$, we do not reject the null hypothesis at 5% level.

Conclusion: There is no significant difference between the correlation coefficients.

Partial correlation

The correlation coefficient earlier discussed measures the association between two characters. We may also consider the simultaneous variation of more than two characters. For example, in coconut shell weight, amount of water, shell thickness, kernel weight, kernel thickness, fruit weight, etc. are known to be correlated with one another such that if we take any set of three characters, they will have correlations among themselves. Fruit weight and nut weight may be strongly positively correlated because a third variable say kernel weight is also strongly positively correlated to both. What if the effect of the kernel weight is eliminated, that is, for all fruits having the same kernel weight, will the association between fruit and nut weight be still strong? A correlation coefficient known as partial correlation coefficient measures the association of two variables after making allowance for their association with other specified variables. The variable whose influence is allowed for in the calculation of the partial correlation coefficient is spoken as the *eliminated* variable. The partial correlation coefficients can be calculated with the help of correlation coefficients by successively accounting for the influences of other variables. The details for the calculation of the partial correlation coefficients are not provided here, but could be obtained from Steel and Torrie (1981).

Regression

The functional relationship of a variable with other variables is often referred to as regression. Unlike correlation, the object of regression analysis is to determine the functional relationship or the equation which relates the variables. With this

function, we are able to explain how much variation in the dependent variable is due to the independent or the *regressor* variable. This function is also used to predict the value of the dependent variable given the value of the independent variable. While the concept of independent and dependent variables is absent in correlation, these variables are clearly defined in regression analysis.

With one independent variable and a dependent variable, the regression is known as simple regression. On the other hand, regression of one dependent variable on two or more independent variables is called multiple regression. If the relationship is linear, it is called linear regression. In many situations, a linear regression model is adequate to describe the relationship of variables. Besides this, linear regression model is easier to interpret and possesses certain mathematical and statistical properties. In fitting regression models, it is advisable to assess adequacy of linear models before venturing into more complex models.

Simple linear regression

When there is one dependent and only one independent variable, and the relationship is assumed to be linear, the regression analysis is called simple linear regression analysis. This involves determining the linear function between the independent (X) and the dependent (Y) variables which is of the form

$$Y_p = \alpha + \beta X + \varepsilon$$

Where, Y_p is the mean of Y for a given value of X; α is the y-intercept or the value of Y_p when X is zero; and β is the regression coefficient which is the change in Y_p for every unit change in X. The random error component is denoted as ε.

The function or line in this case is estimated by obtaining estimates of α and β from the sample. Using the method of least squares, the estimators for α and β are:

$a = \overline{Y} - b\overline{X}$ and $b = \frac{s_{xy}}{s_x^2}$, respectively

Where, $\overline{Y}$ is the mean of Y, $\overline{X}$ is the mean of X, S_{xy} is the sample covariance of X and Y and S_x^2 is the sample variance of X. Substituting these values, the equation of the line is estimated as $\hat{Y} = a + bX$, where $\hat{Y}$ is the predicted value of Y for a given value of $X=x$.

The computation of b can be made simplified by the following formula:

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

Unlike correlation coefficients, where $\rho_{xy} = \rho_{yx}$, the regression of Y on X is different from the regression of X on Y. It should be borne in mind that no attempt should be made to obtain the value of X corresponding to the value of Y from the regression of y on x and vice versa. The two regression lines will intersect at the point $(\bar{X}, \bar{Y})$. Further, if we multiply b and b' we obtain r^2, the square of the correlation coefficient r. The correlation coefficient is therefore the geometric mean of the two linear regression coefficients and is alternatively defined as such. The determination of the independent and dependent variables is not a random process and requires some knowledge of the principles in the process involved.

It may be noted that the method of least squares employed to estimate the regression coefficients involves any assumptions on the population from where the sample is drawn. However, under Gauss-Markov assumptions the procedure offers a grate deal of statistical analysis and inference. These assumptions are:

- The errors (ε) are normally and independently distributed with mean 0 and constant variance;
- The independent variables are non-random and
- No independent variable can be expressed as a linear combination of the remaining independent variables.

The failure of the last assumption is known as *multicollinearity*.

Example

Consider the fruit characteristics of 20 randomly selected West Coast Tall palms shown in Table 5.2. The regression equation of nut weight, NW (Y), on whole fruit weight, FW (X), is determined using the same calculation mentioned in Table 5.3 in connection with the computation of coefficient of correlation.

To determine the equation of the line, we need to estimate α and β. It may be observed that the numerator of the expression for estimating β is the corrected sum of cross products between X and Y and was obtained as 559012.5. The denominator is corrected sum of squares for the variable X and is obtained as 1137923 (from Table 5.3). Thus,

$$b_{yx} = \frac{559012.5}{1137923} = 0.491257$$

$$a = \frac{12090}{20} - (0.491257)\left(\frac{20779}{20}\right) = 94.10854$$

The regression equation of nut weight (Y) on fruit weight (X) is then

$$y = 94.10854 + 0.491257x$$

On the other hand, the regression of fruit weight on nut weight is

$b' = b_{x/y} = 1.7586$ and $a' = -1222.57$

It can be verified that $b_{x/y}\, b_{y/x} = (0.4913)(1.7586) = 0.8640$, which is equal to

$r^2_{FW,NW} = 0.9295^2 = 0.8640$

Selection of the regression line

For a given pair of variables, we can construct two regression lines as indicated above. The choice of appropriate line is based on the purpose of the analysis. For example, if we are interested to know the nut weight without removing the husk of fruit, we may fit a regression of NW on FW. This regression coefficient obtained in this case will also provide us the incremental change in nut weight for every unit change in fruit weight. On the other hand if one wants to know the incremental changes in fruit weight per unit change in nut weight, the regression of FW on NW needed to be fitted.

Test of hypothesis about the regression coefficient

Consider observations on two or more variables taken on n randomly selected units of a sample. Assuming that a linear relation exists between the variables.

Hypothesis: $H_o : \beta = 0$, against $H_1 : \beta \neq 0$

Computation:

$$t_c = \frac{b}{s.e.(b)}$$

where b is the estimate of the regression coefficient, and

$$s.e.(b) = \frac{s_{y/x}}{s_x}\sqrt{\frac{1}{n-1}} \quad \text{and} \quad s_{y/x} = \sqrt{\frac{(n-1)\left(s_y^2 - b s_{xy}\right)}{n-2}}$$

which under the null hypothesis follows the t-distribution with n-2 df

The simplified expression for $s^2_{y/x}$ is of interest to us:

$$s^2_{y/x} = \left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right) - b\left(\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}\right)$$

It may be observed that $s^2_{y/x}$ has two components: The first component is the total sum of squares of variable y with (n-1) degrees of freedom and the second expression is the sum of squares due to regression of y on x with 1 degree of freedom. Therefore, $s^2_{y/x}$ is referred as the residual sum of squares and have df = n – 2.

The aforesaid partitioning of total sum of squares lead to an alternative test statistic, which is based on the *F*-distribution and is obtained as the ratio of the mean squares due to regression to the mean squares due to deviations from regression. Mean squares are obtained by dividing the sum of squares with their corresponding df. The above mentioned ratio has F distribution with df 1 and (n-2) under the null hypothesis.

Decision: Reject H_0 if $t_c > t_{\alpha,\ n-2}$, otherwise, fail to reject H_0.

Example

Using the previous example on 20 randomly selected West Coast Tall palms, test the hypothesis that the nut weight is linearly dependent on the whole fruit weight.

Hypothesis: $H_0 : \beta_{yx} = 0$ against $H_1 : \beta_{yx} \neq 0$

Computation:

Calculate S^2_x, S^2_y, S_{sy}, $\overline{x}$ and $\overline{y}$ as follows:

$$s_x^2 = \frac{(1216^2 + 1445^2 + ... + 1183^2) - \frac{(1216 + 1445 + ... + 1183)^2}{20}}{19} = 59890.68$$

$$s_y^2 = \frac{(662^2 + 735^2 + ... + 555^2) - \frac{(662 + 735 + ... + 555)^2}{20}}{19} = 16730.47$$

$$s_{xy} = \frac{(1216)(662) + 1445)(735) + ... + (1183)(555) - \frac{(20779)(12090)}{20}}{19} = 29421.71$$

$$\overline{X} = \frac{(1216^2 + 1445^2 + ... + 1183^2)}{20} = 1038.95$$

$$\overline{Y} = \frac{(662 + 735 + ... + 555)}{20} = 604.5$$

Using the F-test,

$$F = \frac{b(n-1)s_{xy}}{S^2_{y.x}} = \frac{(0.4913)\ (19)\ (29421.71)}{2403.347367} = 114.2651$$

From the above computations for S^2_y (16730.47), S_{xy} (29421.71), S^2_x (59890.68) and b_{yx} = 0.4913, obtain $S_{y/x}$ as follows:

$$s_{y/x} = \sqrt{\frac{(20-1)\ (16730.47 - (0.4913)\ (29421.71))}{20-2}} = 49.0239$$

$$s.e.(b) = \frac{49.0239}{244.73}\sqrt{\frac{1}{20-1}} = 0.0460$$

Using the t-statistic, $t_c = \frac{0.4913}{0.0460} = 10.689$

Decision: From tables we obtain $t_{(0.05,\ 18)}$ = 2.101; and $F_{(0.05,\ 1,\ 18)}$ = 4.41. These values are less than the corresponding test statistics. Hence, we reject the null hypothesis.

Conclusion: The regression coefficient of nut weight on fruit weight is significantly different from zero. In other words, nut weight is linearly dependent on the fruit weight. The practical utility of this relationship is that based on the values of weight of fruit, the weight of nut may be predicted.

Multiple linear regression

When there are more than one or *p* independent variables in the linear regression model, we refer to it as a multiple (linear) regression model

$$Y_p = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

which can be expressed in matrix notation as:

$$\boldsymbol{Y = X\beta + \varepsilon}$$

Where, $\boldsymbol{Y}$, is a n x 1 vector of observations on the dependent variables, $\boldsymbol{X}$ is a *n* x (*p*+1) matrix of independent variables, $\boldsymbol{\beta}$ is (*p*+1)x1 vector of regression coefficients and $\boldsymbol{\varepsilon}$ is n x 1 vector of unobservable errors. Unlike simple regression, the coefficients here do not fully describe the relation of dependent variable on specific independent variable and hence is distinguished by the term *partial regression coefficient*. Hence β_1 is the partial regression coefficient which is the change in $\boldsymbol{Y}$

for every unit change in x_1 given that x_2, x_3 ....., x_p are held constant. So is the estimation of coefficients. The procedure for estimating the regression coefficients for a multiple linear regression analysis is given below. This step-by-step procedure can be done in MS Excel or any spreadsheet software that can handle matrix operations.

Step 1
Arrange data in matrix format with *n* rows corresponding to the observational units and columns as the variables (refer the data on fruit characteristics used for illustration of correlation coefficients shown in Table 5.2). It is conventional to keep the dependent variable (Y) in the last column. Also add a new column with all values equal to 1 and denote this variable as x_o. This is to include the coefficient α in the regression model. We assume there are $\boldsymbol{p}$ independent variables. Therefore we have to estimate $\boldsymbol{p}$ partial regression coefficients and a constant. Thus the total number of parameters to be estimated becomes *p+1*. The data under *p* independent variables and under x_o is represented by the matrix $\boldsymbol{X}$ of order *n x (p+1)* (i.e. the matrix $\mathbf{X}$ has *n* rows and *p*+1 columns). The data under the dependent variable is denoted by the *n x 1* vector **y**. Similarly, construct the $\boldsymbol{\beta}$ the *(p+1) x 1* vector of parameters and $\boldsymbol{\varepsilon}$ the *n x 1* vector of errors.

Step 2
Obtain the sums of squares and sums of cross products of all *p*+1 independent variables (including x_o) and arrange them accordingly as a matrix. This could be done individually for the columns or carry out the matrix multiplication operation $\boldsymbol{X'X}$ and denote the resulting matrix as $\boldsymbol{S}$. Obtain the cross products of each of the independent variables with the dependent variable or carry out the matrix multiplication $\boldsymbol{X'Y}$ and denote this as column vector **b**.

Step 3
Obtain the inverse of $\boldsymbol{S}$, denoted as $\boldsymbol{S}^{-1}$. The computation of inverse of a matrix is not explained here; it is advised to do it in some spreadsheet software, for example MS EXCEL.

Step 4
The solution for $\boldsymbol{\beta}$ or the vector of estimates of regression coefficients is obtained as $\boldsymbol{S^{-1}b}$; in which the first value is the estimate of $\boldsymbol{\alpha}$, second value is the estimate of partial regression coefficient of first independent variable and so on.

Step 5
Test H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ against H_1: At least one regression coefficient is not equal to zero.

To compute the value of the test statistic, the following sum of squares (SS) need to be worked out:

1. Obtain the total sum of squares for Y (*SSY*)

$$\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Where, Y_i is the observed value of the dependent variable on the i^{th} unit.

2. Obtain the regression SS as follows: Let the estimates of regression coefficients be arranged in $\hat{\boldsymbol{\beta}}$ so that its first element is the constant (estimate of α), second element is estimate of β_1, etc. Then regression SS is the product of corresponding elements of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{b}$.

3. SS due to estimates of partial regression coefficients (β_1, β_2 …) is then obtained by subtracting $(\Sigma y_i)^2/n$ from the regression SS.

4. Residual SS is then obtained by subtracting the SS in (3) from the total SS.

Step 6
Calculation of test statistic: The df's of the SS in 1 to 4 above is respectively *n-1*, *p+1*, *p*, *n-p-1*; where, *p* is the number of independent variables. Obtain the mean SS by dividing the SS with corresponding df. The test statistic is obtained as:

$$F = \frac{\text{Mean Square due to } \beta_1, \ldots, \beta_p}{\text{Mean Square Residual}}$$

which follows the *F*-distribution with parameters *p* and *n-p-1*

Step 7
The R^2 measures the adequacy of fit of the regression model to the observations and is obtained as:

$$R^2 = \frac{\text{Sum of Squares due to Regression}}{\text{SSY}}$$

It is the square of the multiple correlation coefficient which is the correlation between Y and $\hat{\boldsymbol{Y}}$.

Step 8
To test the significance of individual partial regression coefficients, the standard errors of each are obtained as:

Hypothesis: $H_0 : \beta_i = 0$, against $H_1 : \beta_i \neq 0$

Computation:

$$t_c = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}$$

where, $\hat{\beta}_i$ is the estimator of βi, and s.e. $(\hat{\beta}_i)$ is the standard error of $\hat{\beta}_i$ which is obtained by

$$s.e.(\hat{\beta}_i) = \left(\sqrt{MSE}\right)\left(\sqrt{d_{ii}}\right)$$

Where, d_{ii} is the diagonal element of the $\boldsymbol{S^{-1}}$ in the order of β_i.

Decision: Reject H_0 if $t_c > t_{0.05,\ n-p-1}$

Example

Consider the data on fruit characteristics shown in Table 5.2. It is desired to predict the copra weight (CW) based on fruit weight (FW), nut weight (NW), volume of cavity (VC) and kernel weight (KW).

Step 1

Construct the $\boldsymbol{X}$ matrix with 20 rows and 5 columns corresponding to the 4 independent variables, FW, NW, VC and KW with a column of 1's as the first column.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1216 & 662 & \cdots & 172 \\ 1 & 1445 & 735 & \cdots & 187 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1183 & 555 & \cdots & 164 \end{bmatrix}$$

Step 2

Obtain the matrix $\boldsymbol{S = X'X}$ or the matrix of sum of squares and sum of cross products of the variables X_o, FW, NW, VC and KW. (*This can be done using the MMULT function in MS Excel*). Note that the elements of the matrix were already worked out at the time of computation of the correlation matrix as presented in Table 5.3, except for the first row and first column. The sum of products of the aforesaid variables with the dependent variable (CW) denoted by **b** is also shown in the following Tables 5.5a to 5.5c (Steps 2 to 4 involved in the fitting of multiple linear regression equation).

$$S = X'X = \begin{bmatrix} 20 & 20779 & 12090 & 3610 & 6323 \\ 20779 & 22726265 & 13119918 & 3971650 & 6815670 \\ 12090 & 13119918 & 7626284 & 2322420 & 3963032 \\ 3610 & 3971650 & 2322420 & 724300 & 1204150 \\ 6323 & 6815670 & 3963032 & 120415 & 2066707 \end{bmatrix}$$

$$b' = X'Y = \begin{bmatrix} 3680 \\ 3933024 \\ 2293940 \\ 697610 \\ 1194426 \end{bmatrix}$$

Table 5.5a. Results of matrix S and column vector b computation (Step 2)

	Elements of the SSSP matrix (S)					Elements
	X_o	FW	NW	VC	KW	of b
X_o	20	20779	12090	3610	6323	3680
FW	20779	22726265	13119918	3971650	6815670	3933024
NW	12090	13119918	7626284	2322420	3963032	2293940
VC	3610	3971650	2322420	724300	1204150	697610
KW	6323	6815670	3963032	1204150	2066707	1194426

Step 3

Obtain S^{-1} (use *MINVERSE function of MS Excel*)

$$S^{-1} = \begin{bmatrix} 3.007291857 & 0.002436509 & -0.007827522 & 0.014360095 & -0.010592955 \\ 0.002436509 & 1.05056E\text{-}05 & -2.88087E\text{-}05 & 2.46732E\text{-}05 & -1.23334E\text{-}06 \\ -0.007827522 & -2.88087E\text{-}05 & 0.000132055 & -0.000102679 & -7.44443E\text{-}05 \\ 0.014360095 & 2.46732E\text{-}05 & -0.000102679 & 0.000150786 & -1.62634E\text{-}05 \\ -0.010592955 & -1.23334E\text{-}06 & -7.44443E\text{-}05 & -1.62634E\text{-}05 & 0.000189187 \end{bmatrix}$$

Table 5.5b. Results of S^{-1} computation (Step 3)

	X_o	FW	NW	VC	KW
X_o	3.0072919	0.0024365	-0.0078275	0.0143601	-0.0105930
FW	0.0024365	0.0000105	-0.0000288	0.0000247	-0.0000012
NW	-0.0078275	-0.0000288	0.0001321	-0.0001027	-0.0000744
VC	0.0143601	0.0000247	-0.0001027	0.0001508	-0.0000163
KW	-0.0105930	-0.0000012	-0.0000744	-0.0000163	0.0001892

Step 4:
Obtain $\hat{\beta}$

$$\hat{\beta} = S^{-1}b = \begin{bmatrix} 59.0606 \\ -0.0613 \\ 0.2683 \\ 0.1101 \\ 0.0207 \end{bmatrix}$$

Table 5.5c. Estimate of regression coefficients (Step 4)

Estimate of	Calculation	Estimate
Constant (a)	3.0072919 x 3680 + 0.0024365 x 3933024 + ... -0.0105930 x 1194426	59.060623
b_{FW}	0.0024365 x 3680 + ... - 0.0000012 x 1194426	-0.061295
b_{NW}	-0.0078275 x 3680 +..... - 0.0000744x 1194426	0.268304
b_{VC}	0.0143601x 3680 + ... - 0.0000163x 1194426	0.110142
b_{KW}	-0.0105930 x 3680 + ... + 0.0001892x 1194426	0.020722

Step 5
Hypothesis: H_o : $\beta_{FW} = \beta_{NW} = \beta_{VC} = \beta_{KW} = 0$ against H_1 : At least one regression coefficient not equal to zero.

Computation:

1. Total SS (of Y) = 694866 - 3680 x 3680/20 = 17746
2. The regression SS
   = 59.060623 x 3680 + -0.061295 x 3933024 + … + 0.020722 x 1194426
   = 693329.323
3. The SS due to estimates of partial regression coefficients (β_{FW}, β_{NW}, β_{VC}, β_{KW}) is then 693329.323 – 3680 x 3680/20 = 16209.32274
4. Residual SS = 17746 - 16209.32274 = 1536.677261

In the above example the number of independent variables k = 4

Mean SS due to partial regression coefficients = 16209.32274/4
= 4052.330684

Mean residual SS = 1536.677261/(20-4-1)
= 102.44515

F = 4052.330684/102.44515
= 39.5561

Decision: Since $F_c > F_{(5\%, 4,15)} = 3.06$, we reject H_0.

Conclusion: At least one partial regression coefficient is significantly different from zero.

Step 6

Coefficient of determination $R^2 = 16209.32274/17746 = 0.9134$. This implies that the regression equation explains 91.34% of the variation in copra weight.

The individual tests for the parameters are given below (Table 5.5d). Since $t_{o(5\%,15)} = 2.131$, only α and β_{NW} are significantly different from zero.

Table 5.5d. Testing significance of individual coefficients (Step 7)

Coefficients	Estimate	Diagonal element of S^{-1}	Standard error	Test statistic
Constant (α)	59.060623	3.0072919	17.55227819	3.3648
β_{FW}	-0.061295	0.0000105	0.032797471	-1.8689
β_{NW}	0.268304	0.0001321	0.116331442	2.3064
β_{VC}	0.110142	0.0001508	0.124292915	0.8861
β_{KW}	0.020722	0.0001892	0.139221487	0.1488

Note: *Many other topics of interest lies in the purview of regression analysis such as selection of variables to be included in the regression equation, analysis of residuals, step wise regression procedures, etc., but are not discussed here. [Interested readers may refer to text books on Regression Analysis e.g., Draper and Smith (1981)].*

Path coefficient analysis

In biological sciences, the observations for a character may be the outcome of influences of several factors. For example, the copra yield of coconut is influenced by several factors namely number of bunches, the number of nuts per bunch and size of the nuts among others. These characters are in turn influenced by factors such as the rainfall, irrigation levels, fertilizer levels, plant density, etc. If the relationship between the cause and the effect is well defined, it is possible to represent the whole system by a path-diagram which gives an idea of the interrelationships and possible roles of the variables in the system under study. Let us consider the above mentioned example of the copra yield 'Y' of a coconut palm as the result of various causal factors like, number of nut bunches per palm (X_1), number of nuts per bunch (X_2) and nut size (X_3). Let us also assume that these factors are associated as shown in Fig. 5.1.

In Fig. 5.1, copra yield (Y) is depicted as the result of components X_1, X_2, X_3, and some unidentified factors designated by 'R'. Further, we also know that X_1, X_2 and X_3 are correlated among themselves. In this cause-effect relationship we try to ascribe the variability in the effect to contributions from different causal factors. For this purpose, we use the path-coefficients as the measures to quantify the influences of the causal factors on the effect. In Fig. 5.1, a, b, c, and h are the path-coefficients due to the respective variables X_1, X_2, X_3 and R.

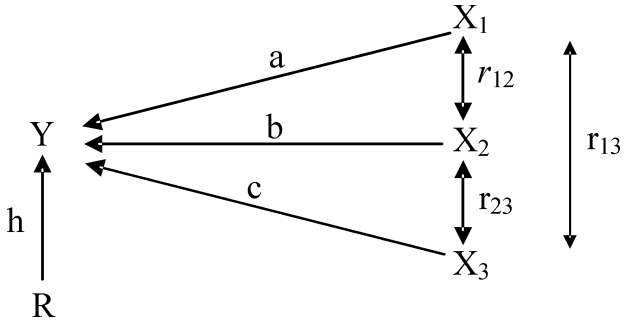


Figure 5.1. Causes and effect relationship.

A path-coefficient is defined as the ratio of the standard deviation of the effect due to a given cause to the total standard deviation of the effect i.e., if Y is the effect and X_1 is the cause, then the path-coefficient for the path from cause X_1 to the effect Y is σ_{x1}/σ_y.

It may be noted that the evaluation of causal models requires clear definition of cause and effect relationships. The estimation of path-coefficients is similar to estimation of regression coefficients but an incorrect theoretical model may result to false and misleading conclusions.

With respect to estimation, the path-coefficients are equivalent to standardized regression coefficients. Contrary to regression analysis, where only single dependent variable is expressed in terms of other independent variables, no such restriction exists for path-coefficient analysis. In other words, path-coefficient analysis may use results of more than one regression analysis.

Assumptions: Relationship among the variables is linear and additive; in case more than one error term is involved, they are uncorrelated with each other; and only one-way causal flows in the system.

Data: Observations on two or more variables on *n* units of a randomly selected sample.

Step 1
Prepare the data following the same format as described for multiple linear regression.

Step 2
Obtain the mean and standard deviation of all variables. Transform the variables to standard random variables with mean zero and variance 1. This is achieved by subtracting the mean from every observation of a variable and then dividing by the standard deviation of that variable.

Step 3
Obtain the correlation matrix and the regression coefficients of the dependent variable on other independent variables.

Step 4
Construct the table of direct and indirect effects in a matrix form. The diagonal elements represent the direct effects, which are nothing but the regression coefficients as obtained above. Unlike correlation, this matrix is not symmetrical. In other words, the indirect effect of the variable X_1 through X_2 is not the same as that of X_2 through X_1. Denoting the standardized regression coefficients of X_1, X_2 on the dependent variable Y as b_1 and b_2, and the correlation coefficient between X_1 and X_2 by r_{12}, the indirect effect of X_1 through X_2 (on Y) is defined as $b_2\ r_{12}$. The indirect effect of X_2 through X_1 is defined as $b_1 r_{12}$.

Step 5
The 'residual effect' or 'residual path term' is defined as $\sqrt{1 - R^2}$, where R^2 is the coefficient of determination of the regression equation defined.

Example

Consider the data on fruit characteristics shown in Table 5.2 and determine the magnitudes of the direct and indirect effects of the fruit characters viz., fruit weight (FW), nut weight (NW), volume of cavity (VC) and kernel weight (KW) on copra weight (CW).

Step 1
Obtain the sample mean and standard deviation (already obtained while computing the correlation coefficients - refer Table 5.3). For example, for FW, the sample mean is obtained as $20779/20 = 1038.95$ and the standard deviation is $\sqrt{1137923/19} = 244.7257$.

Step 2
Transform the data under each variable to standard form by subtracting the corresponding sample mean from the observations and dividing by the standard error. For example, the value 1216 of FW becomes $(1216 - 1038.95)/244.7257 = 0.72$; the value 1445 becomes $(1445 - 1038.95)/244.7257 = 1.66$ and so on. The standardized values are shown in Table 5.6.

Step 3
The correlations between the variables were already worked out in Table 5.4 and will not be changed by the above transformation. Next, work out the partial regression coefficients of FW, NW, VC, and KW on CW using the standardized data shown above. The regression coefficients obtained are as follows:

$b_{FW} = -0.491;\ b_{NW} = 1.136;\ b_{VC} = 0.223;\ b_{KW} = 0.040$

Step 4

Obtain the matrix of direct and indirect effects. Examples of the computation on indirect effects of FW on CW are as follows:

Direct effect = b_{FW} = -0.491

Indirect effect of FW via NW = b_{NW} $r_{FW, NW}$ = 1.136 x 0.929 = 1.055344

Indirect effect of FW via VC = b_{VC} $r_{FW, VC}$ = 0.223 x 0.769 = 0.171487

Indirect effect of FW via KW = b_{KW} $r_{FW, KW}$ = 0.040 x 0.888 = 0.03552

Table 5.6. Standardized scores of variables indicated in Table 5.2

FW	NW	VC	KW	CW
0.72	0.44	-0.01	0.50	-0.39
1.66	1.01	0.32	1.12	0.10
-1.03	-1.07	-1.14	-0.91	-0.88
-1.04	-1.06	-1.14	-0.74	-1.05
-1.18	-1.09	-0.98	-0.91	-0.95
-0.14	0.26	0.15	-0.19	0.33
-0.82	-0.77	-0.65	-0.62	-0.46
-0.60	-0.34	-0.01	-0.87	-0.62
-0.08	0.07	0.15	0.08	0.46
-0.73	-0.92	-0.17	-1.07	-0.85
0.09	0.75	0.80	0.77	1.31
-0.45	-0.27	-0.01	-0.19	0.33
2.16	2.09	2.09	1.89	2.00
1.72	2.04	1.93	1.64	2.16
0.42	0.63	1.45	1.17	0.82
0.54	0.91	0.80	1.40	0.72
-0.32	-0.44	-0.65	-0.69	-0.72
-1.34	-1.29	-0.98	-1.28	-1.31
-0.15	-0.56	-0.82	-0.61	-0.33
0.59	-0.38	-1.14	-0.51	-0.65

Verification

The sum of all the effects will be equal to the correlation coefficient of FW with CW (i.e. 0.772, as can be seen from the correlation matrix given above).

Sum of effects = -0.491 + 1.055344 + 0.171487 + 0.03552 = 0.771351

Similarly other indirect effects may be worked out and presented as follows:

Table 5.7. Direct (diagonal) and indirect effect of copra weight in coconut

Characters	Fruit weight	Nut weight	Volume of cavity	Kernel weight	Sum of effects
Fruit weight	-0.491	1.055	0.171	0.035	0.771
Nut weight	-0.456	1.136	0.206	0.038	0.924
Volume of cavity	-0.377	1.047	0.223	0.036	0.929
Kernel weight	-0.436	1.090	0.200	0.040	0.894

Step 5

It may be verified, as indicated above, that the coefficient of determination of the regression equation considered in this example is $R^2 = 0.9134$. Therefore, the 'residual effect' or 'residual path term' is $\sqrt{(1 - 0.9134)} = 0.294267$.

Alternative approach

The path coefficients can also be worked out from the correlations as described below:

The first step is the formation of equations based on the path–diagram. These equations provide information on the direct and indirect contribution of these causal factors to the effect.

The theoretical basis of these equations may be explained considering the correlation between X_1 and Y, i.e. in Fig. 5.1. Assuming that:

$$Y = X_1 + X_2 + X_3 + R \quad (1)$$

Thus,

$$\overline{Y} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \overline{R}$$

Further, we know that $r_{X_1 Y} = \dfrac{\sigma_{X_1 Y}}{\sigma_{X_1}\sigma_Y}$ $\quad r_{X_1 Y} = \dfrac{\mathrm{cov}(X_1, X_1 + X_2 + X_3 + R)}{\sigma_{X_1}\sigma_Y}$ (2)

By substituting the value of Y in the above equation, we get

which simplifies to

$$r_{X_1 Y} = \frac{\sigma^2_{X_1}}{\sigma_{X_1}\sigma_Y} + \frac{r_{X_1 X_2}\sigma_{X_1}\sigma_{X_2}}{\sigma_{X_1}\sigma_Y} + \frac{r_{X_1 X_3}\sigma_{X_1}\sigma_{X_3}}{\sigma_{X_1}\sigma_Y}$$

$$= \frac{\sigma_{X_1}}{\sigma_Y} + \frac{r_{X_1 X_2}\sigma_{X_2}}{\sigma_Y} + \frac{r_{X_1 X_3}\sigma_{X_3}}{\sigma_Y}$$

Where,

cov (X_1, X_1) = variance of X_1 = $\sigma^2_{X_1}$

cov (X_1, R) = 0 as assumed earlier in the diagram

cov (X_1, X_2) = $r_{X_1 X_2}\sigma_{X_1}\sigma_{X_2}$ as per definition of correlation coefficient

Further, as derived earlier for the diagram above,

σ_{X_1}/σ_Y = 'a', the path-coefficient from X_1 to Y

σ_{X_2}/σ_Y = 'b', the path-coefficient from X_2 to Y,

σ_{X_3}/σ_Y = 'c', the path-coefficient from X_3 to Y.

Thus, equation (2) becomes

$$r_{X_1Y} = a + r_{X_1X_2}b + r_{X_1X_3}c \quad (3)$$

From the above equation (3) it is obvious that the correlation between X_1 and Y can be partitioned into three parts namely:

1. Due to the direct effect of X_1 on Y which amounts to 'a';
2. Due to indirect effect of X_1 on Y via X_2 which amounts to $r_{X1X2}b$; and
3. Due to indirect effect of X_1 on Y via X_3 which equals to $r_{X1X3}c$.

Similarly, we can work out the equations for r_{X_2Y}, r_{X_3Y} and r_{RY}. We thus finally get a set of simultaneous equations as given below:

$$r_{X_1Y} = a + r_{X_1X_2}b + r_{X_1X_3}c$$

$$r_{X_2Y} = r_{X_1X_2}a + b + r_{X_2X_3}c$$

$$r_{X_3Y} = r_{X_1X_3}a + r_{X_2X_3}b + c$$

$$r_{RY} = h$$

Considering only the three defined factors, i.e. X_1, X_2, and X_3, the first three simultaneous equations may be solved to get the values of the path-coefficients viz., a, b and c.

Example

Based on the correlation between copra weight and other fruit characters (Table 5.4), the simultaneous equations can be written as follows:

$$0.772 = b_{FW} + 0.929\, b_{NW} + 0.769\, b_{VC} + 0.888\, b_{KW}$$
$$0.924 = 0.929\, b_{FW} + b_{NW} + 0.922\, b_{VC} + 0.960\, b_{KW}$$
$$0.929 = 0.769\, b_{FW} + 0.922\, b_{NW} + b_{VC} + 0.896\, b_{KW}$$
$$0.894 = 0.888\, b_{FW} + 0.960\, b_{NW} + 0.896\, b_{VC} + b_{KW}$$

The aforesaid simultaneous equations can be conveniently written in matrix notation as:

$$\mathbf{r} = \mathbf{C}\ \mathbf{b}$$

Where, ***r*** represents the column vector of correlation coefficients shown in the right hand side of the above equations; **C** is the matrix of correlation coefficients between the characters FW, NW, VC, and KW; and **b** is the column vector of direct effects.

Now $\mathbf{b} = \mathbf{C}^{-1}\ \mathbf{r}$

Where, $\mathbf{C}^{-1}$ is the inverse of the matrix **C**.

The computations could be done using MS Excel. The $\mathbf{C}^{-1}$ is obtained as:

$$\mathbf{C}^{-1} = \begin{bmatrix} 11.66004 & -16.7724 & 6.835857 & -0.37755 \\ -16.7724 & 41.07369 & -15.148 & -10.9642 \\ 6.835857 & -15.148 & 10.74588 & -1.15645 \\ -0.37755 & -10.9642 & -1.15645 & 12.8971 \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} -0.483 \\ 1.129 \\ 0.230 \\ 0.033 \end{bmatrix}$$

The direct effects obtained in this approach are not exactly the same as in Table 5.7. This is because of numerical error in calculation of inverse of a matrix, etc**.** More accurate values will be the one obtained in the previous method (*i.e.* the standardized regression approach).

Note: *The residual effect can be obtained as:*

$$R_{R,CW} = \sqrt{1 - (b_{FW}\, r_{CW,FW}) - (b_{NW}\, r_{CW,NW}) - (b_{VC}\, r_{CW,VC}) - (b_{KW}\, r_{CW,KW})}$$

$$= \sqrt{1 - (-0.483)(0.772) - (1.129)(0.924) - (0.230)(0.929) - (0.033)(0.894)}$$

$$= \sqrt{1 - 0.913443} = 0.294205$$

The 'residual effect' or 'residual path term' is therefore 0.294205; it may be recalled here that the residual effect in the previous method was 0.294267.

Reference

Draper, N. and Smith, H. 1981. Applied Regression Analysis (2nd edition). John Wiley and Sons, New York. 709p.

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research, 6th edition. Longman, Edinburgh. 146p.

Steel, R.G.D. and Torrie, J.H. 1981. Principles and Procedures of Statistics. McGraw-Hill, Singapore. 572p.

Chapter 6: Basic principles for planning and conducting coconut field trials

An experiment is a planned inquiry to obtain new facts or to verify the results of previous scientific studies. Comparison of varieties, controlled investigation for testing the efficacy of fertilizers or pesticides, specially designed programmes for progeny testing or selection trials, identifying the best medium for germination of embryos, etc. are examples of experiments.

The resource requirements in terms of land, labour and money for coconut field trials are significantly higher compared to annual crops since coconut occupies larger area and requires 4 to 7 years before it can start producing nuts. Proper planning of coconut field trials is, therefore, a must to optimize the use of resources. The principles of experimentation viz., randomization, replication and reducing error, when followed, will result in optimum use of resources. To meet this, appropriate experimental designs should be selected and followed. In this chapter, we will discuss the general aspects for laying and conducting coconut field trials. The description of commonly used experimental designs is deferred to subsequent chapters.

Types of experiments

The coconut field trials can be broadly categorized as follows:

1. Comparative experiments (e.g. evaluation of varieties, comparison of methods, comparison of newly introduced accessions in a genebank with one or more local check varieties).
2. Factorial experiments involving two or more factors; besides comparison of different levels (categories) of each factor, the interaction between two or more factors is also of interest (e.g. to find out the best combination of N, P and K fertilizers for high yield, to find out the best hormone combination for a tissue culture medium).
3. Experiments with mixtures; essentially a factorial experiment, but for any treatment the quantity of levels of different factors when added will have same quantity (e.g. to find out the optimum number of split application of fixed amount of fertilizer for higher yield).
4. Experiments for fitting the response pattern of multi-factors; here the selection of different levels of factors are made in such a way that the response pattern can be best fitted (i.e. the variance of estimated response is a function of the sum of squares of the corresponding levels of factors). The design to meet this criterion is called second order rotatable response surface design.

The other types of experiments include bioassays (to estimate the potency of the test preparation relative to that of the standard preparation in a pharmacological

trial or entomology experiment), experiments using partial diallel crosses to estimate genetic parameters. We discuss here only the comparative experiments in which the effects of two or more treatments are compared.

Treatment

The material or procedure whose performance/effect is to be estimated or compared is referred to as the treatment. The treatment may be varieties, pesticides, nutrient levels, combination of different levels of factors such as sucrose and charcoal in an *in vitro* experiment. Treatments are decided according to the objectives of the experiment.

Experimental unit

An experimental unit or experimental plot is the unit of material to which a treatment is applied. In a fertilizer experiment, a plot of nine contiguous coconut trees may receive a particular dose of fertilizers. A different dose of fertilizer will be applied to another plot and so on.

Treatment effect

Observation on the experimental unit (plot) or observation on a fraction of the experimental unit is made to measure the effect of a treatment. The average of measurements (response) of a treatment is taken as its effect. The response is usually taken as the values of economically important characters (e.g. the copra yield per palm in an agronomic experiment or time taken for flowering of 50% palms in a plot).

If the effects of two treatments are the same, that is the difference of average response between the two treatments is equal to zero, we say that the two treatments are at par. Otherwise, one treatment may be preferred (or superior) over the other. In practice, even the difference of average response between two identical treatments may not be equal to zero. This is because the response is not solely due to the treatments, but influenced by many external factors. It is not possible to generate data in which the contribution of external factors is identical while comparing any two treatments. In this regard, the data collected from a comparative trial is considered as a 'sample' of observations drawn from the populations defined by the treatments. Hence, it is necessary to follow procedures of statistical test of significance to draw conclusions. The statistical test involved partitioning of the 'total' variability in the data to various sources of variation, known as the analysis of variance, which will be discussed in subsequent chapters.

Experimental error

Variation among experimental units may be seen due to:

1. Inherent variability, and
2. Variation that results from any lack in uniformity in the physical conduct of the experiment.

Experimental error is a measure of variation, which exists among observations on experimental units that are treated alike. Experimental designs are evolved for reducing the experimental error. The error mean square (*which will be discussed later*) obtained from the analysis of variance of the experimental data provides an estimate of experimental error.

Control of experimental error

It is desirable to reduce the experimental error to detect even a small real difference between treatments. The techniques used to reduce experimental error are:

1. Blocking
2. Use of auxiliary variable, and
3. Choice of size and shape of plots.

Blocking

It is obvious that when two treatments are applied to identical plots, the difference of the plot values will be the actual difference between the treatments. However, in practice, such identical plots are seldom available. Under these circumstances, we group the experimental units in such a way that the variation is less within group (of plots) when compared to variation between groups. This kind of grouping of homogeneous experimental units or plots is called *blocking* and groups thus formed are called *blocks*. The experimental units are grouped into blocks based on their earlier responses or on the basis of their characteristics influencing the response.

In field trials, contiguous plots are grouped to form blocks, with respect to the direction of fertility gradient. That is, plots within a block will have almost uniform fertility status. Another way to do blocking is to arrange the units in the decreasing or increasing order of magnitudes of certain characteristic value (e.g. yield). With this, the units that are showing sudden changes in values may be avoided.

Auxiliary observations

If auxiliary observations are available, which are not affected by the treatment effects, variation in treatment response may be corrected for variation in the auxiliary variable at the time of data analysis. This technique is known as analysis of covariance.

Formation of plots

In general, a plot of 'n' observations will have variance reduced to the order 1/n. However, this benefit will become less important when 'n' increases. Hence, it is necessary to workout optimum plot sizes to reduce cost of experimentation. It is desirable to have uniform plot size for all treatments in a trial. Otherwise, different treatment effects will have different standard errors.

Replication

The repetition of the same experimental treatment under an investigation is known as replication. The purposes of replication are:

- Provides an estimate of experimental error;
- Improves the precision of the experiment by reducing standard error of the mean, thereby, increasing the scope of inference of the experimental results, and
- Effectively control or minimise the error variance.

Increasing the replications by r times, increases the precision by $\sqrt{r}$ times due to reduction of the standard error to $(1/\sqrt{r})$ times. It is desirable, in general, not to have merely the bare minimum of two replications to estimate error but a larger number to reap other advantages. However, it can be seen that σ^2/r decreases rapidly when r is increased from two onwards initially, but the benefits become less as r increased further. Therefore, increasing the number of replications beyond a certain limit does not bring returns by way of increased precision commensurate with the additional resources to be spent. The number of replications required for a trial is decided, by taken into account a number of factors such as the experimental design chosen for the trial, precision, inherent variability of experimental units, resources available, etc.

Number of replications

One of the general considerations for determining the number of replications in an experiment is that the error variance (mean squares) is estimated with at least 10 to 12 degrees of freedom. The expression given below calculates the number of replications required to enable us to infer the difference between two treatments if significant at a given level, i.e. when the observed difference exceeds a given percent of the mean. One thumb rule is to fix the percentage to secure economic viability of treatments compared. Suppose this difference is fixed as d% of the overall mean μ. Let the coefficient of variation of observations on experimental units (if more than one observation per plot, the average values) be denoted by C%. Then number of replications (r) required for significant difference between two treatment effects at α significant level is obtained from the relation:

$$t_{\alpha,n} = \frac{d}{C}\sqrt{\frac{r}{2}}$$

where, $t_{(a,\ n)}$ is the critical value of t-distribution at α significant level for 'n' df. In practice α will be fixed in advance as 5%. As an approximate, we can take $t_{(5\%,\ n)}$ as 2, the value corresponding to critical value of the t-distribution at 5% significant level for large samples.

Example

Consider a situation in which it is required to detect treatment difference that exceeds 15% of the mean (i.e. d = 15). Let the coefficient of variation C = 8%. Then the number of replication required is given by the relationship:

$$2 < \frac{15}{8}\sqrt{\frac{r}{2}} \quad \text{or} \quad r > 4 \times 8^2 \times \frac{2}{15^2};\ \text{i.e.,}\ r > 2.27$$

Thus, the minimum number of replications required for detecting the given difference at the 5% level of significance is the next integer greater than 2.27, that is 3.

Note: *The number of replications suggested for coconut trials will be discussed later in this Chapter.*

Critical difference

If s^2 is the error mean square, the standard error of the difference of two treatment means is given by

$$\sqrt{s^2\left(\frac{1}{r_1}+\frac{1}{r_2}\right)}$$

Where, r_1 and r_2 are the replications of treatment 1 and 2, respectively. The ratio of difference of two treatment means to its standard error is distributed as Student's *t*. Consequently, the difference will be significant if the ratio is greater than the value of *t* for error degrees of freedom at, as taken usually, 5% level. In other words, the difference is significant if the difference between means is greater than $t_{0.05}\sqrt{s^2\left(\frac{1}{r_1}+\frac{1}{r_2}\right)}$. This expression is independent of which particular pair of treatments is being compared and any difference larger than the obtained $t_{0.05}$ value is considered significant. It is, therefore, called the *least significant difference* or the *critical difference*.

Randomization

The function of randomization is to avoid bias in the estimate of treatment effects and to provide valid estimate of experimental error and thus ensure the validity of the statistical tests. The randomization is effected by the allotment of treatments at random to the experimental units (plots). However, the experimental design chosen or the layout plan may impose certain restrictions on randomization, which will be described later while discussing various experimental designs. The tables of random numbers or random permutations may be used for randomization (Fisher and Yates 1963).

Example

Consider an experiment having four treatments, T_1, T_2, T_3 and T_4 that are to be randomly allocated among 24 experimental units (say, trees) so that each treatment is replicated six times. The first step is to serially number the trees from 1 to 24. By referring a table of random digits, the treatments can be allocated at random in the following alternative methods:

Method 1

- Locate a starting point in the table (say, second row and fifth column). Then select 24 *three* digit numbers. Suppose we obtained the following numbers: 142, 503, 674, 499, 647, … 737, 969, 277, 233, 231. (*Note that the number of experimental units is of two digit*).
- Rank these numbers from 1 to 24 so that the smallest number (96) is rank 1 and the largest number 969 is rank 24. The ranks obtained in order are shown below:

2	10	16	9	15	21	11	14
18	17	20	1	13	8	6	12
3	22	23	19	24	7	5	4

- Take first 6 ranks (i.e. 2, 10, 16, 9, 15 and 21) and assign treatment T_1 to trees (or plots) having these serial numbers. Plots with serial numbers corresponding to next 6 ranks (i.e. 11, 14, 18, 17, 20 and 1) will be assigned Treatment T_2 and so on.

Method 2

- Take the largest *two digit* number divisible by 24 (*note that number experimental units is of two digit and is 24*); which is 96.
- Locate a position in the random number table (say, second row and fifth column).
- For each two-digit random number (less than or equal to 95) read vertically, record the reminder until we get 24 distinct numbers as shown in Table 6.1. The random numbers obtained and corresponding reminders worked out are given below. It may be seen that, the number 24 was not obtained even after considering the subsequent columns (starting with 60). As all other numbers were obtained, the last one can be taken as 24 in this case.
- Now as in the previous method, assign treatment T_1 to plots of serial number corresponding to the first 6 'reminders' (i.e. 14, 2, 19, 1, 16, 12), etc.

Table 6.1. Selection of random numbers following method 2

Random number	Reminder	Random number	Reminder	Random number	Reminder
14	*14*	41	*17*	35	*11*
50	*2*	26	~~*2*~~	66	~~*18*~~
67	*19*	55	*7*	28	*4*
49	*1*	21	*21*	37	~~*13*~~
64	*16*	87	*15*	47	~~*23*~~
84	*12*	91	~~*19*~~	76	~~*4*~~
53	*5*	73	~~*1*~~	25	~~*1*~~
61	*13*	27	*3*	23	~~*23*~~
71	*23*	23	~~*23*~~	38	~~*14*~~
68	*20*	23	~~*23*~~	31	*7*
77	~~*5*~~	18	*18*	30	*6*
9	*9*	90	~~*18*~~	51	~~*3*~~
56	*8*	10	*10*	67	~~*19*~~

Selection of an experimental design

As discussed earlier, suitable design needs to be employed for conducting any experiment. Based on the principle that grouping of experimental units for homogeneity will lead to reduced experimental error, different types of designs are suggested for conducting experiments. Table 6.2 presents distinct features of various experimental designs.

Analysis of data

The statistical procedures to be employed for analysis of data from experimental design depend on the chosen design as well as the type of experiment (as described in the beginning of this chapter). Majority of the situations where experimental designs used in coconut are for comparison among treatments. The test of hypothesis is then $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (i.e. there is no difference among the treatment means) against the alternative hypothesis H_1 that at least one treatment mean significantly different from the rest. Analogous to the partitioning of total variance of the dependant variable in the regression analysis to that due to regression and residual, we partition the total variance to that due to treatments and residual (error) and employ F-test for testing the hypothesis. Partitioning of the variance to different components is referred as Analysis of Variance (ANOVA). There are few assumptions while employ ANOVA to test the aforesaid hypothesis. They are: (i) the underlying model is additive (i.e. effects of treatment, replication, and error are additive); (ii) experimental errors are random, independently and normally distributed with mean zero and common variance (homogeneity of variances). It is found that ANOVA is a robust procedure for the assumption of homogeneity of variances. It is also a practice to transform the data so as to satisfy the assumptions before attempting ANOVA. The ANOVA appropriate to different designs will be discussed in the subsequent chapters while describing the respective designs.

Table 6.2. Summary features for commonly used experimental designs

Name of design	Description
Completely Randomized Design (CRD)	All the experimental units are considered as a homogeneous group; no blocking is required. Generally used in laboratory experiments.
Randomized (Complete) Block Design (RCBD)	Homogeneous groups of experimental units (blocks) made according to a single criteria (i.e. one source direction of variation among experimental units is controlled) and block size (i.e. number of experimental units or plots per block) equals to number of treatments. Common design for field experiments.
Latin Square Design (LSD)	Controlling (eliminating) two sources of variation; number of treatments and replications are equal.
Split-plot Design	Suitable for factorial experiments where large plots are required for applications of different levels of one factor. An appropriate design is selected for these treatments and then each large plot is split into small plots (sub-plots) and randomly allocated the levels of the other factor (or combination of two or more factors), that require only smaller plots.
Strip-plot Design	Instead of randomly allocating treatments within plots of large size, the levels of strip-factor is imposed in a similar way in all plots within a block. In other words, the block is split and each split receives one level of the strip-factor.
Incomplete Block Design	When there are many treatments, all treatments may not be accommodated in a block, i.e. the number of units in a block is less than the number of treatments. This lead to incomplete blocks or blocks that do not have complete set of treatments.
Balanced Incomplete Block Design (BIBD)	As in RCBD, all treatment differences have equal standard error. Often a large number of experimental units is required for BIBD. There is also a restriction that the number of blocks cannot be less than the number of treatments.
Partially Balanced Incomplete Block Design (PBIBD)	Instead of equal standard error for all treatment comparison in PBIBD, two sets of treatment comparisons involved two different standard errors. The number of replications is less compared to BIBD. This design is also used for selection of sample of crosses in a partial diallel experiment.
Lattice Designs	Special case of PBIBDs; useful when number of treatments $v = n^2$ or $n(n - 1)$, where n is any natural number.
Augmented Designs	For testing new accessions (which are usually limited in units and therefore, can only have less number of replication) with existing or released varieties (checks or local control, which can have any number of replications).
Balanced Treatment Incomplete Block Design or Reinforced BIBD	The test treatment is compared with more than one check; test treatment having more number of replications. Only treatment vs. check comparisons have same standard error, but comparison between checks is not considered.
Row-Column Designs	Controlling two sources of variation (LSD is a special case of row-column design); special cases are lattice square; Youden square, etc.
Nested Block Designs	Variability within a block is controlled.

Note: *The details of these experimental designs (except the last three in table 6.2) are discussed in chapters 7, 8 and 9.*

Considerations for planning of coconut trials

In this section, we discuss some general considerations regarding coconut variety and hybrid trials. The information is based on earlier published results from coconut experiments and has been recommended for coconut research workers as guidelines in conducting uniform breeding trials (Santos *et al.* 1996).

Characterization of accessions in field genebank

When planning for germplasm collection trials we should consider the optimum plot size, replication, planting of guard rows, and planting of control or check.

Population size

The recommended sample size for each population or variety in a coconut field genebank ranges from 72 to 96 palms for a heterogeneous Tall population. Though reduction in variance can be achieved by means of suitable transformation of data at the time of analysis, large population size is recommended from the point of utilization of the population in breeding programmes as well as for the production of seed nuts. Lower sample size could be used for homogeneous Dwarfs, but maintaining the same number is advantageous.

Experimental design

Large size plots (e.g. six rows of five palms) are suggested for evaluation of accessions, as assessment of genetic variability is also an objective of such trial, besides the standard characterization of the accessions. Single-row design should be discarded because it is subject to a lot of errors due to absence of 'guard rows'. As large plot size reduces 'between plot variability' within a block, a simple randomized complete block design can be used with three replications. In case of large number of accessions, and/or reduced plot size, BIBDs may be used. In cases where population size is less, Augmented Block Design is recommended.

Check cultivars or control population

A proper evaluation of cultivars in a coconut collection is conducted in relation to a well-known or standard population used as a control. A Dwarf (D) control should be used for the Dwarf ecotypes while a Tall (T) population should serve as a check for Tall varieties.

The frequency of the control cultivars depends on the experimental design chosen. If it is a randomized complete block design, the control(s) are also to be included as treatment(s) and will appear in all the blocks.

In the case of augmented design, the control will usually have more number of replications than the new or introduced accessions under evaluation. In this case, the growth and development of the entries are compared against the control. At the Marc DELORME Coconut Research Station in Cote d'Ivoire, the Malayan Yellow Dwarf (MYD) is the control for the Dwarfs while the West African Tall (WAT)

is used for the Talls. On the other hand, at the PCA Zamboanga Research Centre, MYD is also used for the Dwarfs while the local cultivar Baybay (BAY) is used as the Tall control for the Talls. The West Coast Tall (WCT) serves as varietal control at Central Plantation Crop Research Institute, India.

Planting density

The optimum planting density for Dwarfs is 180 palms/ha (8 m triangular); for the Tall, a density of 143 palms/ha (9 m triangular) is suggested. This density could be increased up to a maximum of 210 palms/ha (7.5 m triangular) for Dwarf ecotypes with small crowns, or when the land is a limiting factor.

Management conditions for accurate germplasm characterization and evaluation

Any source of heterogeneity, which can increase experimental error and reduce accuracy, must be avoided to allow a better evaluation of the germplasm. This requires efficient management of the field genebank through an effective interdisciplinary collaboration among breeders, agronomists and plant protection specialists.

Labelling and sampling

To facilitate the evaluation, the labelling system by field number, row and rank of the palm in the row should be adopted. The sample palms for the evaluation of the different varieties should have a specific mark. The use of aluminium labels where marks are embossed and the use of copper wire as tying material ensure that the tag is durable for many years in the nursery and in the field. In areas where field workers are not highly qualified, a very simple labelling method is more appropriate. Simple labels are generally preferred since size for studying morphological traits and fruit component analysis (FCA) is large, i.e. 30 palms per variety.

A regular harvesting method (monthly - for the Dwarfs and D x D hybrids; and bimonthly - for D x T and T x T hybrids, and Tall ones) should be followed when collecting yield data. For the fruit component analysis, the frequency is bimonthly which is described in greater detail in Santos *et al.* (1996).

Comparison of populations/hybrids

Optimum population size

When comparing populations or hybrids, the optimum size is the same as in the case of comparing accessions, i.e. 72 to 96 palms per population/hybrid. Production of sufficient number of seedlings is the most important activity in the planning stage. In the case of population or hybrid trials, 24 male parents crossed with a sufficient number of female parents (minimum of 48 palms) within a period of three months would be enough to produce sufficient number of seed nuts (144).

Given a selection rate of about 67% at the seedbed nursery stage, these seed nuts will give 96 seedlings. These estimates may vary depending on growing conditions.

Experimental design

The choice of a design will depends on the objective of the trial, the genetic structure of the test material, the number of entries and homogeneity of the experimental field. Randomized complete blocks, the Latin squares, and balanced incomplete block designs are often used by the coconut breeders. The number of replications could be three or more depending upon the design. Multi-location trials could also be adopted for increasing the number of replications. The suggested planting density for Tall (T) x Tall (T), Dwarf (D) x Dwarf (D) and Tall (T) x Dwarf (D) are 143, 180, and 160 palms/ha, respectively.

Plot size

While deciding plot size, the xenia effect (influence of the pollen genotype on the albumen of the nut) and the high degree of out-crossing, which occurs in most of the Tall and hybrid materials should be considered, which necessitate large plots. In the inter ecotype tests with high heterogeneity, a plot size of 24 palms (4 x 6) is suggested while, for the performance test between half sib families, a total of 16 palms (4 x 4) are adequate.

Guard rows and palms

Two border rows (the first and the last) can serve as 'guards' to protect the experimental area. Within the row, the first and the last palm are used as guards. The guard rows should be planted with the same type of material as for the experimental rows. Palms in guard rows should not be sampled to eliminate the external or border effect in gathering data.

Control or check variety

A proper breeding trial requires the use of a locally adopted cultivar or a tested and released hybrid as a control. The PB 121 (MYD x WAT), cultivated worldwide, could be used as an international check for D x T hybrid yield tests; and the WAT x RIT hybrid for T x T trials; in addition to another local hybrid.

Fertilizer application and disease control

Hybrids express their genetic potential better under optimum nutritional conditions. Therefore, it is very important to monitor the nutritional status of the test materials through foliar analysis. A standard rate of fertilizer recommended to the test site, unless fertilizer is a treatment factor or a variable, should be applied on all the treatments besides plant protection measures. When disease tolerance is the main focus, the natural exposure to infection is of course necessary for screening the test materials.

Labelling

Coconut breeding is a long and tedious work, which requires special care in the recording system to avoid errors that could occur at any stage of the experimentation; from hybrid seed production, nursery management and field planting to evaluation. Accession books giving information on the origin and pedigree of the tested materials, the nursery records and the field designs must be kept. The palms should be labelled following the system specified. Moreover, the sample palms should be specifically marked. A file detailing the status of every palm in the trial (replacement, producer, abnormal, dead, border) should be available and updated every year.

Sampling

The vegetative and reproductive characters are observed using 30 random palms for every hybrid entry. For the yield components, the number of bunches and nuts are counted on each individual palm during every harvest, which is conducted bi-monthly (in the case of Tall, T x T, and D x T hybrids), or monthly (in the case of Dwarfs and D x D hybrids). For the fruit component analysis, one nut per palm will be taken and pooled samples analyzed on a per plot basis.

Data analysis

Several data files are required for the efficient management of data in coconut breeding programs *viz.*,

1. Information on the palm's status;
2. Origin and identity of the combinations tested in each trial;
3. Number of bunches and nuts;
4. Copra and oil content;
5. Data on fruit component; and
6. Other qualitative characters.

Agronomic trials

Agronomic trials are conducted to find out optimum production techniques for a locally adopted cultivar or a tested/released hybrid. Compared to breeding trials, the experimental material is expected to have less variation. However, it is important to use coconut palms of uniform age and yield for agronomic trials.

Experimental design

According to the objective, appropriate design has to be chosen, refer to Table 6.2.

Plot size

The optimum plot size (refers to number of palms/plot) is summarized in Table 6.3. The size of the plot or number of palms may be modified to suit the shape of the plot.

Table 6.3. Optimum plot size suggested for agronomic trials

Type	Optimum plot size	Country	Reference
Local cultivar	18-20 palms	Sri Lanka	Joachim (1935); Pieris and Salgado (1937)
West Coast Tall; Hybrids	8 palms	India	Nambiar (1986 a, b)
Seedlings	12 seedlings	Philippines	Alforja *et.al* (1978)

Laboratory experiments

The availability of trained human resources (for initiating the cultures as well as the periodic sub-culturing), laboratory space for maintaining cultures, infrastructure for preparation of adequate culture media, etc. are the major factors that should be considered when deciding the size of the experiment.

Experimental design

It is desired in a laboratory trial to control the variation to the extent possible. A simple way is to repeatedly conduct the experiment so that each trial will be a replication (Karun *et al.* 2003). Separate randomization of treatments is to be followed in each trial. The variation due to technicians, equipment, etc. can also be controlled by the use of appropriate design.

Replication size

Whenever the response of interest per experimental unit is binary (e.g. germinated or not), 15 to 20 units are to be grouped to form a replication.

References

Alforja, L.M., Magat, S.S. and Palomar, C.R. 1978. Assessment of plot size for coconut nursery fertilizer experiment. Philippines J. Coconut Studies, 3(3): 15-20.

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research. 6th edition. Longman, Edinburgh. 146p.

Joachim, A.W.R. 1935. A uniformity trial with coconuts. Tro. Agriculturalist, 85: 198-207.

Karun, A., Muralidharan, K., Sajini, K.K. and Parthasarathy, V. 2003. Design and analysis of coconut embryo culture. CORD 19(1): 48-57.

Nambiar, P.T.N. 1986a. Optimum plot size for D x T coconut palms from fertilizer trial yield data. J. Plantn. Crops, 14: 126-129.

Nambiar, P.T.N. 1986b. Optimum plot size for WC Tall palms from fertilizer trial yield data. J. Plantn. Crops, 16(Suppl.): 489-492.

Pieris, W.V.D. and Salgado, M.L.M. 1937. Experimental error in field experiments with coconuts. Trop. Agric. (Trin.), 89: 75-85.

Santos, G.A., Batugal, P.A., Othman, A., Baudouin, L. and Labouisse, J.P. 1996. Manual on standardized research techniques in coconut breeding. IPGRI/ COGENT, Serdang, Malaysia. 46p.

Further Reading

Abeysinghe. 1986. Calibration experiments on perennial crops using covariance analysis. The case of coconuts. Exp. Agric., 22: 353-361.

Abeywardena, V. 1970. The efficiency of pre-experimental yield in the calibration of coconut yields. Ceylon Coconut Q 21: 85-91.

Addleman, S. 1970. Variability of treatments and experimental units in the design and analysis of experiments. J. Amer. Stat. Assoc. 65: 1095-1108.

Cochran, W.G. and Cox, G.M. 1957. Experimental Designs. Wiley, New York.

Cox, D.R. 1958. Planning of Experiments. Wiley, New York.

Daniel, C. and Bonnot, F. 1987. Setting up experiments in oil palm and coconut plantation. III. Statistical considerations. Oleagineux, 42: 185-188.

Fisher, R.A. 1966. The design of experiments, 8th edition. Hafner Pub. Co. Inc. New York.

Mathes, D.T. 1980. A study on when to conclude a long-term fertilizer trial on coconut yield. Ceylon Coconut Q., 31: 127-133.

Mead, R. 1988. The design of experiments: Statistical principles for practical application. Cambridge University Press.

Panse, S.C. 1955. Some considerations in deciding plot size in field trials with trees and bushes. J. Indian Soc. Agric. Statist., 7: 23-26.

Pearce, S.C. 1976. Field experimentation with fruit trees and other perennial plants. Commonwealth Agricl Bureau. Tech. Communications. No. 23(revised).

Saraswathy, P., and K.S. Krishnan. 1989. Field plot technique for experiments with coconut. J. Plantn. Crops, 16(Suppl.): 481-487.

Shrikhande, V.J. 1958. Some considerations in designing experiments on coconut trees. J. Indian Soc. Agric. Statist. 11: 140-156.

Chapter 7: Basic experimental designs for coconut trials

In the previous chapter, a number of experimental designs are proposed principally to reduce or control the experimental error. Among those, the most commonly used designs in agricultural experiments are:

1. Completely Randomized Design (CRD),
2. Randomized Complete Block Design (RCBD), and
3. Latin Square Design (LSD).

These designs are also referred to as the basic designs as other types of modified designs evolved subsequently from them. In this chapter we discuss these basic designs in more detail and examine how the principles of experimentation (i.e. randomization, replication and local control), described in the previous chapter, are satisfied.

Completely Randomized Design (CRD)

The CRD is the simplest type of layout in which treatments are allotted to the units entirely by chance and all the experimental units are assumed to be homogeneous. In other words, the principle of local control of error does not find a place in its layout. Provided that the experimental area is uniform, CRD gives more precision for treatment comparisons than any other design.

The application of CRD is mainly for laboratory studies, pot culture experiments, etc. where conditions are homogenous. In few cases, CRD layouts were found to be employed for field experiments such as comparison of the control measures for stem bleeding disease in coconut, studies on nutritional requirements, etc., where a single palm is taken as an experimental unit. The advantage of CRD is that we can have different number of replications for different treatments although equal number of replications is advantageous to ensure equal precision of estimates of all the treatment effects.

Randomization

In CRD, the treatments are allotted at random to the experimental units as the name suggests. First, serial numbers are assigned to the experimental units. According to the number of replications required for a treatment, experimental units are selected at random and assigned that treatment.

Example

Consider a field trial conducted for finding the best control measure against stem bleeding disease of coconut. The four treatments compared were T1 (Carbendazim 2.5%), T2 (Tridemorph 4%), T3 (Chiseling and coal tar application) and T4 (Control, i.e. no treatment applied). Each treatment was replicated six times.

Observations on pre-treatment yield and post-treatment yield (in the fourth year) were recorded. As an indication of effectiveness of control measures, the percent increase (+) or decrease (-) in yield in the fourth year over pre-treatment yield was determined.

Layout

The first step in laying out the experimental design was the selection of 24 palms (for imposing the four treatments each having six replications). While selecting palms, utmost care should be taken to assure homogeneity for disease intensity. If variation for disease intensity noticed among the selected palms, CRD would not be the ideal design to conduct the experiment. It is also desired to have homogeneity for yield as well as age for the palms selected for the above mentioned trial. After identification of 24 homogeneous palms for the experiment, they were numbered from 1 to 24. Treatment 1 (T1) was then applied to six randomly selected palms, using random number table. Other treatments (T2 to T4) were also applied in the same way using random number table. The resulting layout of the experiment is shown below (Fig. 7.1).

Palm Number	1	2	3	4	5	6	7	8	9	10	11	12
Treatment	T4	T3	T1	T3	T2	T4	T2	T1	T2	T4	T1	T2

Palm Number	13	14	15	16	17	18	19	20	21	22	23	24
Treatment	T3	T4	T4	T2	T1	T3	T1	T2	T3	T1	T4	T3

Figure 7.1. Field layout for coconut trials in CRD.

Model

The statistical model of CRD is:

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

Where,
Y_{ij} is the observation of j^{th} experimental unit of treatment i
μ is the general mean
τ_i is the effect of treatment i
e_{ij} is the residual variation or error

The error e_{ij} is assumed to be normally distributed with mean 0 and variance σ^2, i.e., $e_{ij} \sim N(0, \sigma^2)$.

Analysis

The observations from a CRD experiment consisting of 't' treatments replicated 'r' times can be arranged in a format as shown in Table 7.1. Unequal number of replications (i.e. treatment T_1 has r_1 replication, treatment T_i has r_i replication, etc.) for treatments will not make any difference in the preparation of the table. With

regard to the experiment on control measures for stem bleeding disease, Table 7.2 shows the data on percentage difference in yield in the fourth year of experimentation over pre-treatment yield.

Table 7.1. Tabulation of data from CRD experiment

Treatments	Replications						Treatment total
1	Y_{11}	Y_{12}	....	Y_{1j}	....	Y_{1r}	T_1
2	Y_{21}	Y_{22}	....	Y_{2j}	....	Y_{2r}	T_2
....	....	....	....	....	....	....	....
i	Y_{i1}	Y_{i2}	....	Y_{ij}	....	Y_{ir}	T_i
....	....	....	....	....	....	....	....
t	Y_{t1}	Y_{t2}	....	Y_{tj}	....	Y_{tr}	T_t
							Grant total $G = \Sigma\Sigma Y_{ij} = (= \Sigma T_i)$

The estimate of the variance for the different sources of variation is obtained from the analysis of variance (ANOVA) table as given in Table 7.3. Since the experimental units are assumed to be homogeneous, only two sources of variation in the data viz., between treatments and within treatments (i.e. error) were considered.

Table 7.2. Effect of stem bleeding disease control treatments as percent increase (+) or decrease (-) in yield over pre-treatment yield

Treatments	Replications						Treatment total	Treatment mean
	RI	RII	RIII	RIV	RV	RVI		
T1	30.3	28.6	26.6	33.4	34.4	29.7	183.0	30.5
T2	37.0	34.7	41.5	36.5	38.1	35.9	223.7	37.3
T3	-10.2	-5.3	-13.3	-6.8	-18.1	-22.1	-75.8	-12.6
T4	-45.3	-19.8	-9.6	-28.9	-49.6	-35.1	-188.3	-31.4
Grand Total							**142.6**	

Table 7.3. Analysis of variance (ANOVA) for CRD

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value
Treatment	(t-1)	Treatment SS	Treatment MS = Treatment SS/(t-1)	$F_{treatment}$ = Treatment MS/ Error MS
Error	(N-t)	Error SS	Error MS = Error SS/(N-t)	
Total	(N-1)	Total SS		

The calculations of DF, SS and MS for the different sources of variation are described and illustrated below:

Computational Procedures

Degrees of freedom (DF)

The degrees of freedom for a source of variability is 1 less the number of levels of the source. Hence, the total df is 1 less than the total number of observations (N-1); the treatment df, one less than the number of treatments (t-1) and experimental

units within treatments, one less than the number of experimental units assigned to a treatment, (ri-1) summed across all treatments,

$\sum_{i=1}^{t}(r_i - 1)$ or N-t.

DF for Total = N-1, where N = $\sum r_i$ = Total number of observations in the experiment (r_i is the number of replications of i^{th} treatment)

DF for Treatment = t-1, where t = number of treatments

DF for Error = N-t

With regard to the above example, N = 24 and t = 4

DF for Total = 24 - 1 = 23,

DF for Treatment = 4 - 1 = 3, and

DF for Error = 24 - 4 = 20.

Sum of squares (SS)

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

where CF = (Grand Total)^2/N and is called the Correction Factor

= $(142.6)^2/24$

Total SS = $[(30.3)^2 + (28.6)^2 + \ldots + (-49.6)^2 + (-35.1)^2] - (142.6)^2/24$

= 22227.58 - 847.28

= 21380.30

Treatment SS = $\sum (T_i^2/r_i)$ - CF

= $(183.0)^2/6 + (223.7)^2/6 + (-75.8)^2/6 + (-188.3)^2/6) - (142.6)^2/24$

= 20788.87 - 847.28

= 19941.59

Error SS = Total SS – Treatment SS

= 21380.30 - 19941.59

= 1438.71

Mean sum of squares (MS)

The mean sum of squares (MS), corresponding to the different sources of variation are obtained by dividing the sums of squares with the associated degrees of freedom

(DF), as indicated below. The error means sum of squares is used as an estimate of the variance.

Treatment MS = Treatment SS/(t-1)
= 19941.59/3
= 6647.196

Error MS = Error SS/(N-t)
= 1438.71/20
= 71.935

Null hypothesis and test of significance

Null hypothesis, denoted by H_0 is the statement, which the researcher wants to disprove by experimentation. In terms of the model, the Null Hypothesis considers that all treatments are equal. In the above example, the null hypothesis is that the percentage change in yield is equal in all the four treatments under study. In the ANOVA, this is tested against the alternative hypothesis that at least in one treatment, the percentage change is different from the rest.

The tests of significance are carried out using F-test based on the ratio of the mean squares. The error mean square is used as the denominator for this F-statistic and the numerator will be the mean squares due to the source whose effects are to be tested for their significance.

$F_{treatment}$ = Treatment MS/Error MS
= 6647.196/71.935
= 92.405

After these calculations, you may arrange these values in an ANOVA table, as presented in Table 7.4.

Table 7.4. ANOVA for treatment effect depicted as percent change in yield

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated* (0.05)
Treatment	3	19941.59	6647.196	92.405	3.10
Error	20	1438.71	71.935		
Total	23	21380.30			

*Tabulated value for $F_{0.05}$ for numerator DF = 3 and denominator DF = 20 is 3.10

From the ANOVA table we observe that the calculated F value (92.405) for the treatments is higher than the tabular F value of 3.10 ($F_{0.05}$ with 3 and 20 DF). Hence, the treatment differences are significant, i.e. the null hypothesis is rejected. It is not implied here that all the treatment differences are significant but we are sure that at least one treatment is different from the rest.

Next, we compare the significance of individual treatments by calculating the treatment means and comparing their differences for significance. The difference between observed treatment means is the estimate of difference between the corresponding population means.

The difference between two treatment means is considered as significant if it exceeds the critical difference (CD), at the required level of significance. The formula of CD (this is more popularly known as LSD or Least Significant Difference) for comparison of two treatments i and j is:

$$CD = t_{\alpha, N-t} \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

Where, r_i and r_j are the number of replications for treatments *i* and *j'*, respectively. Usually the level of significance is taken as 5%.

In the example, the CD (at 5%) for comparison of any two treatments (both are having equal number of replications 6) is obtained as:

$$CD = t_{0.05,20} \sqrt{71.9355 \times \frac{2}{6}} = 2.086 \times 4.8968 = 10.21$$

From Table 7.2, it may be seen that the largest difference was observed for treatment T4 (control) followed by T3 (chiselling and coal tar application). These two treatments are significantly different as the difference between their respective means (18.8) is greater than the CD (10.21). Improvement in yield was observed with the two chemical application treatments (T1 and T2). Even though the increase in yield is higher in T2 (37.3%), it is not statistically significant from T1 (30.5%).

Conclusion

Since the F-test calculated value is greater than the tabular F value we conclude that at least one treatment is different. To identify which treatments are different, comparison of means was done and revealed that yield reduction is significantly larger in the control than in the rest of the treatments. Improvement in yield was observed with the two chemical control measures.

Randomized Complete Block Design (RCBD)

In many situations, the knowledge of the researcher regarding the experimental material enables him/her to group the units in relatively homogeneous groups or blocks, each equal in size to the number of treatments, before allotting the treatments. This often helps in reducing the errors affecting the treatment comparisons. The resulting design is called the Randomized Block Design (RBD) or Randomized Complete Block Design (RCBD). It is the simplest experimental design that employs

all the three basic principles of experimentation and perhaps the most commonly used design in agricultural and biological investigations.

This design may be used when the experimental units are heterogenous but can be grouped in such a way that the number of units in a group (block) is equal to the number of treatments. This allows the complete set of treatments to be present in each of the blocks thus the design, complete block. The objective of forming blocks is to apportion the total variation of experimental units in such a manner as to render the variation among the blocks as large as possible and thereby reduce the variation among the units within a block to the maximum extent. Accordingly, the characteristics chosen as criteria for grouping the experimental units into blocks are those expected to be associated with the measure of the effect of the treatments. This characteristic could be a qualitative or discrete such as vertical description of stem, overall appearance/shape of crown, colour of petiole, leaf spiral direction, etc., or a continuous one such as length of petiole, length of leaf bearing position, length of spikelet, weight of fruit, weight of kernel, nut yield per palm, etc. In the former case, the blocks are formed easily by grouping the units belonging to the same class according to the chosen characteristic. In the latter case, the units are arranged according to (descending or ascending) order of magnitude of the characteristic and blocks of successive units equal in number to the treatments are formed. In most of the field trials, contiguous plots with uniform fertility gradient (as far as possible) form the block.

Randomization

In each block the treatments are allotted once each to the units at random. Since in each of the blocks, all the treatments occur exactly once, the blocks are considered complete. For randomly allocating the treatments, the plots will be conveniently numbered first and then the treatments. Treatments are then assigned to the plots at random within each block. For every block, separate randomization is to be carried out.

Example

Consider an evaluation trial of nine coconut cultivars (AOT, AGT, PHOT, FMS, SSG, FJT, CCT, JGT, LCT). The comparisons were made for average annual yield based on four consecutive years.

Layout

Based on the slope and/or fertility gradient, the experimental area could be divided to form three blocks. Blocking is done perpendicular to the direction of the variation (i.e. the direction of fertility gradient) so that plots within a block will have similar soil/fertility characteristics. Note that in this example, the plots are the experimental units and each plot consists of nine palms. For RCBD, the block size is equal to number of treatments (therefore, each block has a size of nine plots) and number of replications is equal to the number of blocks. It may be noted that the total

number of palms in a block is 81 resulting to a total of 243 experimental palms. In addition, it will also have border palms and the observations from those palms will not be considered for analysis. The following figure (Fig. 7.2) depicts the sample field layout of the cultivar evaluation trial in RCBD (border palms are not shown in the layout).

Block 1								
CCT	LCT	FMS	AGT	FJT	PHOT	AOT	JGT	SSG
Block 2								
FJT	JGT	LCT	PHOT	FMS	AGT	SSG	AOT	CCT
Block 3								
JGT	SSG	CCT	AOT	LCT	PHOT	FJT	FMS	AGT

Figure 7.2. Field layout for coconut hybrid evaluation trial in RCBD.

Model

The statistical model for RCBD is:

$$y_{ij} = \mu + \tau_i + \rho_j + e_{ij}$$

where,
ρ_j denotes the j th replication or block effect and other symbols denote as in the case of CRD.

Analysis

The analysis of RCBD with i treatments replicated j times and arranged in j blocks is rather simple. The observations from a RCBD can be arranged in the form of a two-way table as shown in Table 7.5.

It is important to note that, when a plot (experimental unit) has more than one observation, only the average value is subjected for analysis of variance. For example, in the cultivar evaluation trial mentioned above, the average yield of the nine palms will be taken as the observation from a plot (experimental unit). The data from the trial is shown in Table 7.6.

The estimate of the variance for the different sources of variation is obtained from the analysis of variance (Table 7.7). Compared to CRD ANOVA, one additional source of variation (i.e. between blocks) is included here.

Table 7.5. Data tabulation in RCBD

Treatments	Blocks						Treatment
	1	2	…..	j	….	r	total
1	Y_{11}	Y_{12}	….	Y_{1j}	….	Y_{1r}	T_1
2	Y_{21}	Y_{22}	….	Y_{2j}	….	Y_{2r}	T_2
….	….	….	….	….	….	….	….
i	Y_{i1}	Y_{i2}	….	Y_{ij}	….	Y_{ir}	T_i
….	….	….	….	….	….	….	….
t	Y_{t1}	Y_{t2}	….	Y_{tj}	….	Y_{tr}	T_t
Block total	B_1	B_2	…..	B_j	….	B_r	$G = \sum\sum Y_{ij}$ $(=\sum B_j$ or $\sum T_i)$

Table 7.6. Average number of nuts per palm

Treatments	Replications			Treatment	Treatment
	I	II	III	total	mean
AOT	74.95	54.51	62.60	192.06	64.02
AGT	80.18	71.13	77.80	229.10	76.37
PHOT	70.91	61.45	68.80	201.16	67.05
FMS	65.49	55.63	58.70	179.82	59.94
SSG	93.80	77.65	82.60	254.05	84.68
FJT	69.26	51.01	61.70	181.97	60.66
CCT	90.83	78.75	85.80	255.38	85.13
JGT	71.11	80.13	74.60	225.84	75.28
LCT	120.51	79.80	98.70	299.01	99.67
Block total	737.04	610.05	671.30	2018.39	
Block mean	81.89	67.78	74.59		

Table 7.7. ANOVA for Randomised Complete Block Design (RCBD)

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value
Treatment	(t-1)	Treatment SS	Treatment MS = Treatment SS/(t-1)	$F_{treatment}$ = Treatment MS/ Error MS
Block	(r-1)	Block SS	Block MS = Block SS/(r-1)	F_{block} = Block MS/ Error MS
Error	(t-1)(r-1)	Error SS	Error MS = Error SS/(t-1)(r-1)	
Total	(N-1)	Total SS		

The calculations of DF, SS and MS for the different sources of variation are described and illustrated below:

Degrees of freedom (DF)

DF for Total = N-1, where N = t.r = total number of observations in the experiment
DF for Treatment = t-1, where t = total number of treatments
DF for Blocks = r-1, where r = total number of blocks (i.e. replications)
DF for Error = (t-1)(r-1)

With regard to the above example, t = 9, r = 3; therefore N = 27,
DF for Total = 27 - 1 = 26
DF for Treatment = 9 - 1 = 8
DF for Replication = 3 - 1 = 2
DF for Error = 8 x 2 = 16

Sum of squares (SS)

Total SS = $\sum\sum Y_{ij}^2 - CF$,
Where,
CF = $(\text{Grand Total})^2/N$
Total SS = $(74.95^2 + 54.51^2 + \ldots\ldots + 79.80^2 + 98.70^2) - (2018.39)^2/27$
= 156760.7 - 150885.1
= 5875.602

Treatment SS = $\sum T_i^2/r$ - CF
= $(192.06^2 + 229.10^2 + 201.16^2 + 179.82^2 + 254.05^2 + 181.97^2 + 255.38^2 + 225.84^2 + 299.01^2)/3 - (2018.38)^2/27$
= 465454.4/3 - 150885.1 = 155151.48 - 150885.1
= 4267.73

Block SS = $\sum B_j^2/t - CF$
= $(737.04^2 + 610.05^2 + 671.30^2)/9 - (2018.38)^2/27$
= 1366019.18/9 - 150885.1
= 897.643

Error SS = Total SS – Treatment SS – Block SS
= 5875.602 - 4267.73 - 897.643
= 710.228

Mean squares (MS)

The mean squares (MS) corresponding to the different sources of variation are obtained by dividing the sums of squares by the associated degrees of freedom (DF), as indicated below. The error mean squares is used as an estimate of the variance.

Treatment MS = Treatment SS/(t-1)
= 4267.73/8 = 533.466

Block MS = Block SS/(r-1)
= 897.643/2 = 448.821

Error MS = Error SS/(t-1) (r-1)
= 710.228/16 = 44.389

Null hypothesis and test of significance

For the aforesaid experiment, the null hypothesis is that the average number of nuts per palm for the nine coconut cultivars is the same. This null hypothesis is tested against the alternative hypothesis that at least for one cultivar the average number of nuts per palm is different from the rest. In terms of the model, the null hypothesis will be that τ_I 's are all equal.

The test of significance is carried out using F-test based on the ratio of the mean squares. The error mean square is its denominator and the mean squares due to the source whose effects are to be tested for their significance is the numerator.

$F_{treatment}$ = Treatment MS/Error MS
= 533.466/44.389
= 12.018

F_{block} = Block MS/Error MS
= 448.821/44.389
= 10.111

Arrange the above calculated values in the ANOVA table, as given in Table 7.8.

Table 7.8. ANOVA of nuts per palm

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated*
Treatment (cultivars)	8	4267.730	533.466	12.018	2.59
Block	2	897.643	448.821	10.111	
Error	16	710.228	44.389		
Total	26	5875.601			

*Tabulated value for $F_{0.05}$ with 8 and 16 degrees of freedom

Since calculated F value (12.023) for the treatments is higher than the tabular F value of 2.59 ($F_{0.05}$ with 8 and 16 DF), the null hypothesis is rejected. This implies that the yield of at least one of the cultivars is different from the rest. To find which cultivars are different from each other, comparison of means should be done.

CD (at significant level α for comparison of any two treatments is given by the formula:

$$CD = t_{\alpha,(t-1)\times(r-1)}\sqrt{\left(\frac{2}{r}\right)MSE}$$

In the example, tabulated t value is 2.12 (corresponding 16 DF at 5% level). Then,

$$CD = 2.12\sqrt{\frac{2\times 44.84}{3}} = 11.53$$

From Table 7.6, it can be seen that the cultivar FMS has the lowest yield (59.94) which is significantly lower than the cultivars yielding more than 59.94+11.53 = 71.47 i.e. yield of FMS is significantly lower than the yield of JGT, AGT, SSG, CCT, and LCT. Note that the yield of LCT is significantly higher than the yield of all the other cultivars. To simplify the presentation of results, we arrange the treatment means in ascending order and form (overlapping) groups of treatments that are not significantly different. It can be seen from the following that the cultivars FMS, FJT, AOT and PHOT are not significantly different. Similarly the AOT, PHOT and JGT are on par and so on.

FMS	FJT	AOT	PHOT	JGT	AGT	SSG	CCT	LCT
59.94	60.66	64.02	67.05	75.28	76.37	84.68	85.13	99.67

The cultivars that are underlined together have insignificant yield differences. This can be shown in another format using superscript, as shown below:

FMS[a]
FJT[a]
AOT[ab]
PHOT[abc]
JGT [bcd]
AGT [cd]
SSG [d]
CCT [d]
LCT [e]

The varieties with same alphabet as superscript are not significantly different from each other.

Conclusion

Since the F-test is significant, we conclude that the treatment effects are significantly different. Comparison of treatment means revealed that LCT yielded significantly higher than the rest of the cultivars. The lowest yield was from FMS, which is not significantly different from the yields of FJT, AOT and PHOT.

Latin Square Design (LSD)

We have seen that the RCBD is intended to reduce error with respect of one (set of) factor(s) used for forming the blocks of experimental units like slope of land,

fertility gradient, etc. When there are two (sets of) factors contributing to the heterogeneity of experimental units, the experimental units are formed into groups differently for the two (sets of) factors, such that ignoring one blocking factor, the other would give a blocking system as in the case of RCBD. In other words, we will be dealing with a two-way source of variation that needs to be taken care of while forming the blocks. For example, in field trials there could be a fertility gradient in two directions – both parallel and at right angles to the ploughed rows. This may be attributed to the usual blocking and the other due to the residual effects of the treatments applied earlier.

In the presence of two sources of variation, the label/serial number of the experimental units can be arranged in rows and columns. Blocks to take care of the variation in one direction constitute the rows and blocks to take care of the other direction of variation constitute the columns. If blocks formed in either direction are complete (i.e. all treatments occur only once in a block), the arrangement will be a square consisting of equal number (say, 't') of rows and columns. Under such situation, by adopting a *Latin Square Design* (LSD) we can remove the variability in both the sources of variation. A *Latin square* is an arrangement of 't' symbols in rows and columns such that every symbol occurs only once in each row and once in each column. On replacing the 't' symbols in a *Latin square* with the treatments, we get a Latin Square Design. The variability due to differences in rows (one source of blocking factor) as well as columns (another source of blocking factor) can be removed from the error SS.

The application of LSD in coconut field trials is limited as it requires large number of replications and identification of exact direction of two sources of variation in the field is difficult. Since the two sources of variations are removed, the available DF for error will also be less making it less attractive for field trials. Hence, LSD is not appropriate when there are several number of treatments and the experimental area and test materials are limited.

Randomization

The randomization procedure for LSD is not straight forward. First step requires selection of a standard Latin square of appropriate order, say t. A Latin square, with its first row and first column elements in alphabetic order, is said to be in a '*standard form*'. Sets of Latin squares in '*standard form*' for different integers are provided in *Statistical Tables* by Fisher and Yates (1963) and also the method of selection of a square at random.

For the purpose of randomization, separately number the rows and columns of the selected Latin square. Keeping the first row unaltered, rearrange the remaining rows at random. Similarly, the columns are rearranged at random. To complete the randomization, assign treatments at random to the letters of the Latin square.

Example

Consider a laboratory experiment in which germination of zygotic embryos of four

coconut cultivars is tested. There are only two technicians to inoculate embryos. A person can inoculate 20 embryos in the morning and 20 in the afternoon, i.e. 20 embryos each of the four cultivars can be inoculated in a day. Incidentally, a 'plot size' of 20 is good enough to estimate the percentage germination. By using a 4 x 4 Latin square arrangement, the variation due to technicians and time of inoculation (this constitute variation of one direction) and variation between days (the second source of variation) can be eliminated.

Note: *It may be noted here that the adoption of LSD for this experiment may not be the most appropriate as the DF for error is very little.*

Layout

The Latin square of order 4 will have symbols A, B, C and D which were randomly allocated to the treatments as A:WCT, B:PHOT, C:WAT and D:LCT. Hence, the Latin square in the standard form (in treatment symbols) is as follows (Fig. 7.3):

WCT	PHOT	WAT	LCT
PHOT	WAT	LCT	WCT
WAT	LCT	WCT	PHOT
LCT	WCT	PHOT	WAT

Figure 7.3. Chosen Latin Square in standard form.

It may be seen that each treatment is appearing only once in a given row as well as column. The next step is to randomize the arrangement. Randomly arrange all the rows except the first row and then randomly arrange the columns. The layout for LSD after randomization is presented in Table 7.9 along with the percent germination of embryos.

Table 7.9. Percent germination of embryos in trial conducted using LSD

Person	Day 1	Day 2	Day 3	Day 4
Technician-1 AM	**WCT**	**WAT**	**PHOT**	**LCT**
	25	**25**	**80**	**55**
Technician-2 AM	**WAT**	**WCT**	**LCT**	**PHOT**
	10	**40**	**65**	**85**
Technician-1 PM	**PHOT**	**LCT**	**WAT**	**WCT**
	85	**70**	**20**	**30**
Technician-2 PM	**LCT**	**PHOT**	**WCT**	**WAT**
	65	**75**	**45**	**20**

Model

The statistical model for LSD is:

$$y_{ijk} = \mu + \tau_i + \rho_j + \chi_k + e_{ij}$$

where,
ρ_j and χ_k denote the effects of j^{th} row and k^{th} column effects, respectively. Other symbols are as mentioned in the case of CRD.

Analysis

The SS due to rows and columns are obtained similar to the replication SS in RCBD; first treating the rows as blocks and then columns as blocks. The total SS and treatment SS are obtained as in the CRD (or RBD) and the error SS with DF (t-1)(t-2) is obtained by subtracting the SS due to treatments, rows and columns from the total SS. To work out the treatment, Technician (row) and day (column) SS, first sum the relevant observation and arrange as shown in Table 7.10.

Table 7.10. Data summarized for rows, columns and treatments

Rows	Sum	Columns	Sum	Treatments	Sum	Mean
Technician-1 AM	185	Day1	185	WCT	140	35.00
Technician-2 AM	200	Day2	210	WAT	75	18.75
Technician-1 PM	205	Day3	210	PHOT	325	81.25
Technician-2 PM	205	Day4	190	LCT	255	63.75
Grant total	**795**		**795**		**795**	

Correction Factor = $CF = (\text{Grand Total})^2/N$
$= (795)^2/16 = 39501.56$

Total SS = $\sum\sum Y_{ij}^2 - CF$
$= 25^2 + 25^2 + \ldots + 65^2 - 39501.56 = 49525 - 39501.56$
$= 10023.44$

Row (Technician) SS = $\sum R_i^2/r - CF$
$= (185^2 + \ldots + 205^2)/4 - 39501.56 = 39568.75 - 39501.56$
$= 67.19$

Column (day) SS = $\sum C_k^2/r - CF$
$= (185^2 + \ldots + 290^2)/4 - 39501.56 = 39631.25 - 39501.56$
$= 129.69$

Treatment SS = $\sum T_i^2/r - CF$
$= (140^2 + \ldots + 255^2)/4 - 39501.56 = 48968.75 - 39501.56$
$= 9467.19$

Error SS = Total SS – Row SS – Column SS – Treatment SS
$= 10023.44 - 67.19 - 129.69 - 9467.19$
$= 359.37$

From the above calculations, ANOVA is tabulated as shown in Table 7.11.

Table 7.11. ANOVA for percent germinated embryos

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated*
Rows (Technicians)	3	67.19	22.40	0.3739	4.76
Columns (Days)	3	129.69	43.23	0.7217	
Treatments	3	9467.19	3155.73	52.6869	
Error	6	359.37	59.90		
Total	15	10023.44			

*Tabulated value for $F_{0.05}$ with 3 and 6 degrees of freedom

From the ANOVA, it is obvious that only the treatment effects are significantly different. Neither the technicians nor days seem to have influenced the germination of embryos.

The comparison of cultivars for average germination can be made using CD. As t-value for 6 DF at 5% is 2.447, the CD is obtained as 2.447 x √(2 x 59.90/ 4) = 13.39. The treatment means in Table 7.10 can be compared based on this CD.

Conclusion

WAT had the lowest average percentage germination which was significantly less compared to PHOT and LCT. The germination of PHOT is significantly higher than WCT. There was no significant difference in germination between PHOT and LCT.

Repeated Latin Squares

In case two or more Latin squares, say *n* in number, are used in an experiment, the model will additionally include the parameters for effect of square and treatment vs. square. Thus, the parameters in such case will be general mean, treatment effect, effect of square, treatment vs. square interaction, rows within square, columns within square and the error. The total sum of squares will then be accordingly split. The total SS with (st^2-*1*) DF and the treatment SS with (*t-1* DF) are obtained in the usual manner. The SS due to the squares with (*s-1* DF) is then obtained similar to that of treatment SS. The SS due to rows within squares and columns within squares [(each with *s(t-1)* DF)] are obtained by subtracting the SS due to squares, respectively from the rows SS and column SS each with *st-1* DF as there will be a total of st rows/columns, t from each of the s squares. The SS due to treatments within the squares with *s(t-1*) DF is also calculated similarly from the *st* treatment totals taken separately in each of the squares and subtracting the treatments SS from it. The DF associated with the error will be *(s-1)(t-1)(t-2).* The error SS is obtained by subtracting the SS due the squares, treatments, treatments vs. squares, rows within squares and columns within squares from the total SS.

Reference

Fisher, R.A. and Yates, F. 1963. Statistical tables for biological, agricultural and medical research, 6th edition. Longman. 146p.

Chapter 8: Experimental designs for coconut trials with modified blocking

The usefulness of randomized complete block design is restricted to situations where the block size is equal to the number of treatments. To overcome this, incomplete block design was proposed in which the number of units or plots is smaller than the number of treatments. Another situation is where sufficient experimental units may not be available to replicate the treatments as in the case of screening trials of newly introduced germplasm. If an incomplete block design is ensuring equal precisions of the estimates to all pairs of treatment effects, it is called Balanced Incomplete Block Design (BIBD). It was found that the Balanced Incomplete Block Designs require large number of replications which is not possible for certain situations. This led to the development of Partially Balanced Incomplete Block Designs (PBIBD). For germplasm screening trials, the Augmented Block Designs are commonly used. We discuss these designs in this Chapter.

Balanced Incomplete Block Design (BIBD)

Balanced Incomplete Block Design is an arrangement of 'v' treatments in 'b' blocks of equal size 'k' and satisfying the following conditions:

1. Block size is less than the number of treatments;
2. All the treatments in a block are distinct;
3. Each treatment appears in exactly 'r' blocks; and
4. Each pair of distinct treatments appears together in λ blocks.

The parameters necessary for defining a BIBD are therefore v, b, r, k and λ. The following parametric relationships holds good for a BIBD:

Total number of experimental units = $v.r = b.k$
Further the following relationships hold good
$r(k-1) = \lambda(v-1)$, and
$v \leq b$

Because of the above relationships, BIB designs may not be applicable for certain situations. However, for a large number of parametric combinations, the BIBD is available and can be obtained either by the methods of construction as explained in text books (Dey 1986) or using software such as SPBD (Statistical Package for Block Designs) developed by the Indian Agricultural Statistics Research Institute, New Delhi, India. The package also performs the ANOVA for a BIBD.

Randomization

On choosing BIBD arrangement with 'v' symbols, randomly assign the treatments

to the symbols. Then arrange the blocks at random and finally allocate the treatments within each block at random to the respective plots.

Example

Consider an experiment to test the skill of pollinators. There were nine pollinators engaged for the production of hybrid seednuts. Obviously, the yield of a palm influences the overall production of hybrid seednuts. Therefore, it is desirable to group the mother palms as uniformly as possible for annual nut yield. In the present study, the palms were grouped as those yielding between 30 to 35 nuts; 35 to 40 nuts and so on resulting into 12 separate groups. The group size varied between 3 and 6 palms.

Since the group size is less than the number of pollinators, it is not possible to use a RCBD for this experiment. On the other hand, with block size 3 and number of palm groups as 12, we can construct a BIBD involving 12 x 3 = 36 experimental units (palms). The arrangement of the design using the symbols 1, 2, …, 9 is shown in Table 8.1. It may be verified from the table that each symbol is repeated in $r = b.k/v = 12 \times 3/9 = 4$ blocks. Every pair of symbols appear in $\lambda = r.(k-1)/(v-1) = 4 \times 2/8 = 1$ number of blocks.

To obtain the layout of the experiment, first we randomly assign the treatments to the symbols, e.g. symbol 1 to Pollinator-4 (P4), symbol 2 to P1, 3 to P8, 4 to P9, 5 to P2; 6 to P7; 7 to P3; 8 to P5 and 9 to P6. The layout of the experiment (after randomization) is shown in Table 8.1. The response variable is taken as percentage of hybrid seedlings obtained and the values are also shown in the Table 8.1.

Table 8.1. Field layout of the BIBD (data generated) along with parameters (v=9, b=12, r=4, k=3, λ=1)

BIBD arrangement			FIELD LAYOUT				Hybrid seedlings (%)			Block Total (B_j)
			Block	Treatments within block						
1	2	3	1	P7	P5	P4	40	55	65	160
4	5	6	2	P3	P5	P6	72	58	25	155
7	8	9	3	P5	P2	P1	63	58	67	188
1	4	7	4	P9	P3	P4	41	80	61	182
2	5	8	5	P5	P8	P9	52	71	49	172
3	6	9	6	P8	P7	P6	78	46	33	157
1	6	8	7	P8	P3	P2	69	71	61	201
2	4	9	8	P9	P1	P6	38	70	36	144
3	5	7	9	P9	P7	P2	34	41	52	127
1	5	9	10	P2	P4	P6	58	68	41	167
2	6	7	11	P3	P7	P1	74	44	71	189
3	4	8	12	P8	P1	P4	77	61	68	206
				Total						2048

Model

The statistical model for BIBD is same as that of RCBD:

$y_{ij} = \mu + \tau_i + \rho_j + e_{ij}$

where,
Y_{ij} is the observation of of treatment i in j^{th} block
μ is the general mean
τ_i is the effect of treatment i
ρ_j denotes the j^{th} block effect
e_{ij} is the residual variation or error

It may be noted here that unlike in RCBD, data from BIBD is not orthogonal. This is because in BIBD, only a fraction of treatments appears in any block and a difference of two block totals (with different sets of treatments) is therefore not exactly equal to the true difference of those two block effects. In other words, in the case of BIBD, partitioning the total variance into components such as 'treatment', 'block' and 'error' is not as straight forward as in RCBD.

Analysis

The calculations of DF, SS and MS for the different sources of variation are described and illustrated below:

Degrees of freedom (DF)

DF for Total = N-1, where N = v.r = Total number of observations
DF for Treatment = v-1, where v = Total number of treatments
DF for Blocks = b-1, where b = Total number of blocks
DF for Error = (N – 1) – (v – 1) – (b – 1)

With regard to example mentioned in Table 8.1, v = 9, b=12, r = 4, k = 3; therefore, N = 36 and

DF for Total = 36 - 1 = 35
DF for Treatment = 9 - 1 = 8
DF for Blocks = 12 - 1 = 11, and
DF for Error = 35 - 8 - 11 = 16

Sum of squares (SS)

From the table, obtain the total SS and block SS as in the case of RCBD

Total SS $= 40^2 + 55^2 + \ldots + 68^2 - (2048)^2/36$
$= 7673.556$
Block SS $= 160^2/3 + 155^2/3 + \ldots. + (206)^2/3 - (2048)^2/36$
$= 2037.556$

As indicated earlier, the Block SS obtained above cannot be considered as the SS due to blocks exclusively owing to the non-orthogonal nature of data. Since comparison of block effects is not the objective of the experiment, there is no need to obtain an adjusted Block SS. However, it is necessary to obtain adjusted treatment SS which is estimated as shown in Table 8.2.

Table 8.2. Computation of adjusted treatment sum of squares

Treatments	Treatment Total (T_i)	Block numbers in which the i^{th} treatment occurs	Sum of block totals corresponding to the block numbers of previous column* ($B_{j(i)}$)	Adjusted treatment totals $Q_i = T_i - [(B_{j(i)})/k]$
P1	269	3,8,11,12	727	26.667
P2	229	3,7,9,10	683	1.333
P3	297	2,4,7,11	727	54.667
P4	262	1,4,10,12	715	23.667
P5	228	1,2,3,5	675	3.000
P6	135	2,6,8,10	623	-72.667
P7	171	1,6,9,11	633	-40.000
P8	295	5,6,7,12	736	49.667
P9	162	4,5,8,9	625	-46.333
Total	2048			0

*Corresponding to P1, the block totals 188, 144, 189, 206, and so on.

In the table, it may be verified that the sum of Q_i equals to zero, which serve as a check for arithmetic calculations. The Treatment SS (adjusted) is then obtained as:

$$(k/\lambda v)\ \Sigma\ Q_i^{\ 2} = (3/9)\ [26.667^2 + 1.333^2 + \ldots + (-46.333^2)] = 5254.815$$

Null hypothesis and test of significance

In the aforesaid experiment, the null hypothesis is that the percentages of hybrid seedlings produced by nine pollinators are the same. The tests of significance are carried out using F-test based on the ratio of the mean sum of squares and summarized in the ANOVA table as depicted in Table 8.3.

$$F_{treatment} = \text{Treatment MS/Error MS} = 656.8519/23.82407 = 27.57093$$

Table 8.3. ANOVA table for percentage hybrid seedlings produced

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated*
Treatment (adjusted) (Pollinators)	8	5254.815	656.8519	27.57093	2.59
Block (Unadjusted)	11	2037.556	185.2323		
Error	16	381.1852	23.82407		
Total	35	7673.556			

*Tabulated value for $F_{0.05}$ with 8 and 16 degrees of freedom

From the ANOVA table we can observe that the calculated F value (27.57093) for the treatments is higher than the tabular F value of 2.59 ($F_{0.05}$ with 8 and 16 degrees of freedom). Hence, the treatment differences or the percentages of the hybrid seedlings produced by the pollinators are significantly different. In other words, the skill of the tested pollinators varied.

Following this we need to test the significance of difference between the individual pollinators. For this purpose, we need the adjusted treatment means and are obtained as follows:

Adjusted treatment mean = Overall mean + Adjusted treatment effect

Adjusted treatment effect is $k\ Q_i/v\ \lambda$,

Where, Q_i is defined as in Table 8.2.

Accordingly, adjusted P1 mean = 2048/36 + 3 x 26.667/9
= 56.889 + 8.889 = 65.778

The adjusted treatment means are shown in Table 8.4.

The CD for comparison of two treatments is given by the expression

$$CD = t_o\sqrt{\frac{2kE}{\lambda v}}$$

Where, t_o is the tabulated value t for error DF at 5% level and E is the error mean square. The CD for comparison is thus 2.12 x √(2 x 3 x 23.82407/9) = 8.448.

Table 8.4. Adjusted treatment means

Treatments	Adjusted mean
P6	32.667[a]
P9	41.445[b]
P7	43.556[b]
P2	57.333[c]
P5	57.889[c]
P4	64.778[c]
P1	65.778[cd]
P8	73.445[d]
P3	75.111[d]

In Table 8.4, the treatment means bearing common superscripted alphabet are not significantly different. For example, the means of P9 and P7 are having same superscript 'b' and are therefore, on par.

Conclusion

It may be observed that the skills of the pollinators P1, P8 and P3 are superior and at par. The performance of pollinator P6 is the worst while P9 and P7 are also not satisfactory.

Partially Balanced Incomplete Block Designs (PBIBD)

Balanced Incomplete Block Designs are not available for every parametric combination. Also in some cases, the number of replications required when a BIBD used may be very large. These constraints led to the development of Partially Balanced Incomplete Block Designs (PBIBD). While BIBD assures equal variance to all the paired comparison of treatment effects, in the case of PBIBD, there will be two groups of treatment comparisons but within group, the variance will be equal. In other words, when a PBIBD is used, treatment comparisons are not made with equal variance but with either of the two possible variances for treatment comparisons. The formal definition of PBIBD and various methods of construction are beyond the scope of this manual. Nevertheless, certain easily available PBIBD will be described.

One special case of PBIBD is the lattice design that exists when number of treatments is a perfect square (i.e. when there are 2, 4, 9, 16, 25, etc. treatments). Such a design is called square lattice. The construction of square lattice designs with three replications is very easy and illustrated with an example below.

Example

Let the number of treatments $v = 9 = 3^2$. Arrange the numbers 1 to 9 in a 3 x 3 square, in which each row represents a block. Another 3 blocks of the design is obtained by treating the columns as blocks. Next, superimpose a Latin square of order 3 on the 3 x 3 square arrangement and assign to blocks treatments corresponding to the letters. If one more replication is required, a mutually orthogonal Latin square may be superimposed and so on. The steps involved in the construction are shown in Table 8.5.

Table 8.5. Construction of square lattice design (v=9; k=3; r=3)

Block	First replication			Block	Second replication			Latin square: Order-3			Latin square (Super-imposed)			Block	First replication		
I	1	2	3	IV	1	4	7	A	B	C	A1	B2	C3	VII	1	6	8
II	4	5	6	V	2	5	8	B	C	A	B4	C5	A6	VIII	2	4	9
III	7	8	9	VI	3	6	9	C	A	B	C7	A8	B9	IX	3	5	7

Analysis

The analysis of variance of data from PBIBD is similar to BIBD but the calculation of critical difference is slightly different. As mentioned earlier, in a PBIBD, there will be two CDs for treatment comparisons.

For detailed analysis of data from PBIBD, readers may refer to Dey (1986).

Augmented Block Design

An Augmented Block Design is an essentially Incomplete Block Design in which a portion of the treatments (called check treatments) are arranged as in a standard block design (e.g. RCBD or BIBD) and to those blocks, the remaining treatments having less number of replications (called test treatments) are added.

An Augmented Block Design is useful for screening new treatments such as genotypes, insecticides, herbicides, drugs, etc. The number of new treatments, *n*, may be very large or the experimenter can afford only one or two replications of these new treatments. One such situation is the preliminary evaluation of new crosses or newly collected germplasm. Often the amount of germplasm collected from the exploration trips may not be adequate for a complete replicated trial. Moreover, all the accessions may not be of promising types. Similarly, different hybrid combinations need to be tested for their worthiness before laying down a full-fledged evaluation trial. In these situations, we are interested to compare the new (test) treatments (i.e. the new accessions, new hybrids, etc.) with that of one or more check treatments (i.e. released hybrids, popular cultivars, etc.). In this situation, we are not interested in determining the differences among paired treatments but in testing the worthiness of test treatments over the check treatments. In other words, we are interested only in a part of all possible paired comparisons for which the variance balanced design are not efficient. The recommended design for such situation is Augmented Block Design, which was proposed by Federer (1956).

Randomization

The augmented design follows the standard randomization procedure for the known design in control treatments or check varieties. Test treatments or new varieties are randomly allotted to the remaining experimental units. The different treatment entries are assigned to a block at random with the provision that no treatment appears more than once in a block.

Example

A breeder wants to evaluate the performance of eight new hybrids in comparison with three traditional hybrids (WCT x COD, COD x WCT and LCT x GBGD) and a popular local cultivar (WCT). Thus, in this experiment, there are eight test treatments (t) (denoted by H1, H2, …, H8) and four check treatments (c). The availability of new hybrid seedlings varied between 12 and 20. However, sufficient number of check variety seedlings is available. Keeping in view of future requirement for gap filling, it was decided to have a plot size of nine seedlings per test treatment. In other words, seedlings are just sufficient to have only a single replication for the test treatments. Also note that the breeder is interested only in comparing new hybrids with the checks rather than the comparison among new hybrids or among the checks. In other words, we are not interested in paired comparison among all the 12 (8 + 4) treatments in this experiment.

Layout

The experimental area is divided into three blocks, each block can accommodate seven treatments (with a plot size of 9 seedlings). Therefore, a RCBD was chosen for the check treatments (number of treatments = 4 and number of

replications = 3). Each block we can accommodate 7 - 4 = 3 test treatments, giving the possibility to include 3 x 3 = 9 test treatments in this trial. As we have only 8 test treatments, it was decided to accommodate 3 test treatments each in the first and third blocks and only 2 test treatments in the second block.

For random allocation of these treatments in the experiment:

1. Allocate the four check treatments to each block randomly; and
2. Allocate the eight new hybrids (test treatments) at random to the remaining experimental units (3 each in first and third blocks and 2 in the second block).

The layout of the experiment is shown in Fig. 8.1. During the first year of planning, it was decided to generate some data on drought tolerance of the new hybrids and therefore observations were made on epicutical wax content (microgram/ cm^2). The average values were shown along with the treatments in Fig. 8.1. The procedure of analysis is adopted from Federer (1956 and 1961).

I	H8	LCT x GBGD	WCT	H3	WCT x COD	COD x WCT	H7
	(74)	(78)	(78)	(70)	(83)	(77)	(75)
II	WCT	COD x WCT	WCT x COD	LCT x GBGD	H1	H5	Non-experimental plot
	(91)	(81)	(79)	(81)	(79)	(78)	
III	H4	LCT x GBGD	WCT x COD	H2	WCT	COD x WCT	H6
	(96)	(87)	(92)	(89)	(81)	(79)	(82)

Figure 8.1. Layout of an augmented block design along with observations on epicutical wax content (microgram/cm^2).

Model

The statistical model for the Augmented Block Design is the same as that of a Randomized Block Design.

$$y_{ij} = \mu + \tau_i + \rho_j + e_{ij}$$

Where,

Y_{ij} is the observation of treatment i in j^{th} block
μ is the general mean
τ_i is the effect of treatment i
ρ_j denotes the j^{th} block effect
e_{ij} is the residual variation or error

Analysis

The analysis is carried out as follows:

Step1

Obtain the block totals, check treatment totals and means, grand total, grand total of check treatments and block sum of the test treatments. For this, arrange the data as shown in Tables 8.6a and b.

Table 8.6a. Block-wise arrangement of check treatments' values

Check treatments	Blocks			Total	Mean
	I	II	III		
WCT x COD	83	79	92	254	84.67
COD x WCT	77	81	79	237	79.00
LCT X GBGD	78	81	87	246	82.00
WCT	78	91	81	250	83.33
Total	316	332	339	987	329.00

Table 8.6b. Block-wise arrangement of test treatments' values

Blocks	Test treatments (Hybrids)								Sum of test treatments in blocks	Sum of all treatments in blocks
	H1	H2	H3	H4	H5	H6	H7	H8		
I			70				75	74	219	535
II	79				78				157	489
III		89		96		82			267	606

Step 2

Obtain the total SS and Block SS (unadjusted) as follows:

Grand total = 535 + 489 + 606
= 1630

Total SS = $74^2 + 78^2 + \ldots\ldots + 79^2 + 82^2 - (1630)^2/20$
= 807.00

Block SS = $535^2/7 + 489^2/6 + 606^2/7 - (1630)^2/20$
= 360.0714

Treatment SS = $(254^2+\ldots+250^2)/3 + (79^2+89^2+75^2+74^2) - (1630)^2/20$
= 575.6667

Step 3

Obtain the adjusted treatment SS which is computed based on adjusted block effects, adjusted overall mean, adjusted check-treatment effects and adjusted test-treatment effects.

Calculation of adjusted block effects (B_j):

B_j = [j^{th} block total - means of all check treatments – values of test treatments appeared in the j^{th} block]/Number of check treatments

B_1 = [535 – (84.67 + 79.00 + 82.00 + 83.33) – 70 – 75 – 74]/4
= [535 – 329 – 219]/4 = -3.25

B_2 = (489 – 329 – 157)/4 = 3/4 = 0.75
B_3 = (606 – 329 – 267)/4 = 10/4 = 2.50

Note: *The sum of block effects added to zero, which may be used for verifying the calculations.*

For calculations of adjusted overall mean (m) follow as described below:

Denoting the number of replications of i^{th} check treatment as r_i and number of test-treatments as t_i

m = {grand total – Σ [(r_i – 1) x i^{th} check treatment mean]
- Σ [t_i x i^{th} adjusted block effect]}/total number of treatments

= {1630 – [2 x 84.67+…+2 x 83.33] – [3 x (-3.25) + 2 x 0.75 + 3 x 2.50]}/12
= {1630 – 658 – (-0.75)}/12 = 81.0625

Calculation of adjusted check-treatment effects (C_i):

C_i = i^{th} check treatment mean - adjusted overall mean

C_1 = 84.67 – 81.0625 = 3.6042
C_2 = 79.00 – 81.0625 = -2.0625
C_3 = 82.00 – 81.0625 = 0.9375, and
C_4 = 83.33 – 81.0625 = 2.2708

Calculation of adjusted test treatment effects (i.e. H_k):

H_k = k^{th} test treatment value - corresponding block effect - m

H_1 = 79 - 0.75 - 81.0625 = -2.8125
H_2 = 89 - 2.50 - 81.0625 = 5.4375
H_3 = 70 - (-3.25) - 81.0625 = -7.8125
H_4 = 96 - 2.50 - 81.0625 = 12.4375
H_5 = 78 - 0.75 - 81.0625 = -3.8125
H_6 = 82 - 2.50 - 81.0625 = -1.5625
H_7 = 75 - (-3.25) - 81.0625 = -2.8125
H_8 = 74 - (-3.25) - 81.0625 = -3.8125

As in the case of block effects, the sum of all the treatment effects (i.e. check and test treatment effects) equals to zero. This verifies the accuracy of computations. Based on the above, we obtain the adjusted treatment SS as:

Adj. TSS = m x grand total + Σ (B_j x j^{th} block total) + Σ (C_i x i^{th} check treatment total) + Σ (H_k x k^{th} test treatment value) – Block SS (without subtracting the correction factor)

= 81.0625 x 1630 + [(-3.25)535+...+2.5 x 606] + [(3.6042 x 254) + ... +2.2708 x 250] + [-2.8125 x 79 +...+(-3.8125) x 74] – ($535^2/7$ + $489^2/6$ + $606^2/7$)

= 285.0954

Error SS (SSE) = Total SS - Adj.TSS - Block SS

= 807.00 - 285.0954 - 360.0714 = 161.8332

The SS of check treatments (which are laid out as in a 'standard design', in the present example, as RCBD), is obtained as usual. Denoting the total for i^{th} treatment as T_{Ti}, and number of replications of i^{th} check treatment as $r_{i.}$

Check treatment SS = Σ T_{Ti}^{2}/r_i – Corresponding Correction Factor
(*The Correction Factor mentioned above is based on the check treatments alone*)
Check treatment SS = (254^2+...+250^2)/3 - $(987)^2/12$ = 52.9167

On subtracting check treatment SS from adjusted treatment SS, we get the combined SS due to test treatments as well as check vs. test treatments as:

Adj. test and test vs. check SS = 285.0954 - 52.9167 = 232.1787

The aforesaid SS can be arranged in an ANOVA as shown in Table 8.7a. It may be noted here that the SS due to test vs. check and adjusted test treatment SS are combined in the ANOVA Table 8.7a. It is desired to separate these two SS, but not necessary for making paired comparisons between check treatment and test treatment. The estimate of error variance provided by the ANOVA is used for the calculation of CD for such tests.

Table 8.7a. ANOVA (Part-1) of epicutical wax content

Sources of variation	DF	Sum of squares	Mean sum of squares	F-value	F-Tabulated*
Blocks (ignoring treatments)	2	360.0714	180.04		
Treatments (Adjusted)	11	285.0954	25.92	0.961	4.03
Check treatments	3	52.9167	17.64	0.654	4.76
Test treatments and test vs. check	8	232.1787	29.02	1.076	4.15
Error	6	161.8332	26.97		
Total	19	870.0000			

It is evident from the ANOVA that there were no significant differences among treatment effects. However, for demonstration, we discuss the computation of critical difference for comparing different kinds of treatment effects. There are four types of treatment comparisons in this kind of trials as indicated below:

1. Between test treatment and check treatment
2. Between two check treatments
3. Between two test treatments in the same block
4. Between two test treatments not belonging to the same block

The objective of augmented design is mainly reflected by critical difference mentioned between test treatment and check. Comparison between two check treatments is also important. The computations of these two critical differences are described below.

The critical difference for comparison of test treatment and check treatments are given by the following expression in which MSE is the error mean square, t_o is the t-value against error DF, c is the number check treatments and r is the number of blocks.

$$CD_{test}\ (5\%) = t_o\sqrt{MSE\ [1 + 1/r + 1/c + 1/(c.r)]}$$
$$= 2.447 \times \sqrt{26.9722} \times (1 + 1/3 + 1/4 + 1/12)$$
$$= 2.447 \times 6.70475 = 16.406$$

CD for comparison of two check treatments is,

$$CD_{check}\ (5\%) = t_o\sqrt{(2 \times MSE/r)}$$
$$= 2.447 \times \sqrt{(2 \times 26.9722/3)}$$
$$= 2.447 \times 4.24$$
$$= 10.375$$

It may be noted here that, comparisons are to be made based on the adjusted treatment effects. It is a practice to provide the treatment SS (ignoring blocks) along with the ANOVA of augmented design. The following are the necessary steps for its estimation:

$$\text{Test treatment SS} = 79^2 + 89^2 + \ldots + 75^2 + 74^2 - (79 + 89 + \ldots + 75 + 74)^2/8 = 505.8750$$

Subtracting the test treatment SS and check treatment SS from the (unadjusted) treatment SS, we get SS due to test treatment vs. check treatment as:

Test vs. check SS = 575.6667 - 505.8750 - 52.9167 = 16.875

The SS can be arranged as in Table 8.7b. One may note that there are no change in error SS, etc. in this part. No test of relevance is provided by this second part and may be skipped.

Table 8.7b. ANOVA (Part-2) of epicutical wax content

Sources of variation	DF	Sum of squares	Mean sum of squares	F-value	F-Tabulated*
Treatments (ignoring blocks)	11	575.6667	52.333	1.940	4.03
Check treatments	3	52.9297	17.643	0.654	4.76
Test treatments	7	505.8750	72.268	2.679	4.21
Test treatments vs. controls	1	16.8750	16.875	0.626	5.99
Error	6	161.8332	26.972		

Conclusion

In the example, treatment differences are not significantly different. However, one limitation of the aforesaid experiment is the few DF for error which is an outcome of the chosen RBD for the trial. Had we increased the replication to 4 the DF for error would have been more than 12. Situations of this kind should be foreseen at the time of planning an experiment.

References

Dey, A. 1986. Theory of Block Designs. Wiley Eastern Ltd., New Delhi. 268p.

Federer, W.T. 1956. Augmented (or hoonuiaku) designs. Hawaiian Planter's Record. VSS, P. 191-208.

Federer, W.T. 1961. Augmented designs with one way elimination of heterogeneity, Biometrics. 17: 447-473.

Chapter 9: Experimental designs for multiple factors

Under natural conditions, different factors play concurrent roles in deciding the overall outcome of an experiment. Thus, to study the realistic output of an experiment, we need to consider multiple factors that play a concomitant coherent role rather than the single independent factors. A factor refers to a group of treatments of similar nature. In some experiments, the treatments are constituted by the combination of different levels (categories) of two or more factors. For example, designated treatments for different quantities of water used for irrigation can be grouped and called as 'levels' of the 'irrigation factor'. Similarly, different quantities of fertilizer constitute the levels of the factor 'fertilizer'. Combinations of different levels of irrigation and fertilizer factors then constitute the treatments of a factorial experiment involving these two factors. To illustrate, if irrigation has 3 levels and fertilizer has 2 levels, then the factorial experiment will have a total of 3 x 2 = 6 treatments. For conducting a factorial experiment, an appropriate design say, CRD, RCBD or LSD may be used. In certain situations, the factorial experiments are conducted using experimental units of different sizes depending on the requirement of the factors. The split plot, strip plot designs and any combination or variation thereof are examples of such experiments. Following a general description of factorial experiments, the analysis for split plot and strip plot designs are described in this Chapter.

Factorial experiments

Factorial experiments deal with simultaneous variations in more than one factor. Following traditional procedure, one may investigate the problems one by one, varying a single factor at a time in a simple experiment. The soundness of the approach rests on the assumption that the responses to the factors, in the aforesaid example, viz., different levels of water supply, different amounts of fertilizers, are independent of one another. However, to make such an assumption, under all situations, is not accurate. The factors may not have only independent or additive effects, but may also interact with each other. In order to find from an experiment whether the factors actually interact with each other or are independent in their effects, one is required to investigate the effects of these factors together in one and the same experiment. This will allow comparison of all the possible combinations of the levels of the factors.

The factorial experiment alone can provide information regarding the interactions between various factors under study. The factorial scheme provides comprehensiveness of the conclusions drawn. It might appear that the increase in comprehensiveness is achieved only at the cost of precision of the comparisons relating to the response to individual factors. Far from it, there is an increase in precision due to what is known as hidden replications in the experiment seen when the levels of the other factors are ignored.

Notations and terminologies

It is conventional to denote the name of a factor in upper case letter. For example, the irrigation factor is denoted by 'I'. Similarly, 'F' denotes the factor fertilizer. The levels of a factor is denoted by the numbers 0, 1, 2, etc. as subscript to the symbol of the factor. As an illustration, let the three levels of the irrigation factor be 25, 50 and 75 L/day/palm. Then I_0 denotes the treatment 25 L/day/palm; I_1 denotes the treatment 50 L/day/palm; and I_2 denotes the treatment 75 L/day/palm. Suppose the factor fertilizer also has 3 levels viz., F_0, F_1 and F_2, then there will be 3 x 3 = 9 treatments and the experiment is referred to as a 3^2 factorial. If there are 4 factors each with 2 levels, we call it as 2^4 factorial. If all the factors in an experiment are of equal number of levels, it is called symmetrical factorial experiment while it is called asymmetrical factorial experiment if factors are of different levels.

Keeping the levels of other factor(s) constant, the simple effect of a factor or the difference of observation for two levels of a factor could be determined. If Y_{uv} is the observation of the treatment combination I_uF_v, then the simple effects of the factor irrigation are:

$Y_{10} - Y_{00}$, $Y_{11} - Y_{01}$, $Y_{20} - Y_{10,}$ etc.

The 'main effect' of a factor is the average of its simple effects over a variety of conditions arising from the replications as well as the repetition of a particular level of a factor with different levels of other factor(s).

When simple effects of a factor differ for different levels of the other factor(s), there is said to be an interaction between the factors. Interaction between 2 factors is called first order interactions or 2-factor interaction; interaction involving 3 factors is called second order interaction or 3-factor interaction, and so on.

When block designs are used for laying out a factorial experiment, one can see that the block size becomes very large when the number of levels and/or number of factors increases. For example, an experiment with 5 factors each at 2 levels required a block size of $2^5 = 32$. A large block size will result to higher variability and hence, it is necessary to go for incomplete blocks. As observations from all the treatments are utilized for defining the main effects as well as for testing of interaction, neither blocks of arbitrary size nor assigning the treatments at random (as in the case of incomplete block design) can be advocated. However, a technique known as confounding is available to overcome this situation. Reduction in block size is achieved by confounding certain interaction effects with the block effects so that they become one and the same.

The construction of confounded factorial arrangements when the number of levels of each factor is same and is a prime or prime power (i.e. number of levels could be 2, 3, 4, 5, 7, 8, 9, etc.) is straight forward when the block size is a power of number of levels. To illustrate, consider a 3^4 experiment. The number of levels

of factor is 3, which is a prime number. For this experiment, confounded arrangements can be constructed when the block size is 3, 3^2, or 3^3. As block effect and the confounding interaction effects become identical, we need to confound interaction(s) having DF equal to block effects. It may be noted here that the treatment sum of squares in a factorial experiment (factors having s levels each) can be split into components having DF (s-1) each.

For example, consider a 3^4 experiment (four factors with 3 levels each). The treatment DF is 3^4-1 = 80 and each of its components will have 3-1 = 2 DF. In other words the treatment SS in this case can be split into 80/2 = 40 components of 2 DF each. The method of constructing confounded factorial experiment use this property to decide upon the number of interaction effects (more correctly the 'components') to be confounded.

If the number of blocks is b, then, the number of components to be confounded is equal to (b-1)/(s-1). In the above example, if block size is 3^2 = 9, it is required to confound (9-1)/(3-1) = 4 components. Since higher order interactions are difficult to interpret, their components are generally confounded. Different methods of construction of confounded designs are presented by Cochran and Cox (1957) and Das and Giri (1986).

The confounding can be *complete*, in the sense that no information will be available on the confounded interaction or it can be *partial* so as to provide information on the confounded interaction but with less precision. Where only a fraction of the total factorial combinations are used, it is known as *fractional factorials* which can provide information on the main effects and lower order interactions (Cochran and Cox 1957; Das and Giri 1986). These are useful, especially for initial screening purposes, involving several factors at a time and/or the higher order interactions, i.e. interactions involving many of the factors at the same time are considered to be negligible or absent.

Example (Simple factorial experiment)

Consider an experiment to study simultaneously the effect of irrigation and fertilizer on coconut yield. The experiment is a 3^2 factorial with irrigation levels as mentioned earlier. The three fertilizer levels representing half, full and one and a half quantity of fertilizers recommended for coconut are denoted as F_0, F_1 and F_2, respectively. The treatment combinations can be represented as I_0F_0; I_0F_1; I_0F_2; I_1F_0; I_1F_1; I_1F_2; I_2F_0; I_2F_1 and I_2F_2. It is also a practice to represent the treatments by dropping the symbols and denoting the factor names as *00, 01, 02, 10, 11, 12, 20, 21* and *22*.

Layout

The experiment was conducted in RCBD with three replications. Each treatment plot has nine palms and border rows were provided in all the sides of the plots to reduce the border effects of treatments. The layout of the experiment is shown in Fig. 9.1.

B1		I_1F_2 60		I_2F_0 55		I_0F_2 64		I_1F_0 63		I_0F_0 53		I_2F_2 71		I_1F_1 66		I_2F_1 65		I_0F_1 64	
B2		I_1F_1 58		I_1F_2 71		I_0F_2 58		I_2F_2 65		I_2F_0 53		I_0F_1 54		I_0F_0 52		I_2F_1 58		I_1F_0 59	
B3		I_2F_2 67		I_0F_2 61		I_2F_0 60		I_1F_1 67		I_1F_2 68		I_0F_0 45		I_2F_1 58		I_1F_0 63		I_0F_1 59	

Figure 9.1. Field layout for factorial experiment along with plot means.

Model

The statistical model for a factorial experiment in RCBD is:

$$Y_{(uv)j} = \mu + \rho_j + i_u + f_v + (if)_{uv} + e_{(uv)j}$$

Where,

$Y_{(uv)j}$ is the observation of treatment combination I_uF_v in j^{th} block

μ is the general mean

ρ_j denotes the j^{th} block effect

i_u is the main effect of u^{th} level of irrigation

f_v is the main effect of v^{th} level of fertilizer

$(if)_{uv}$ is the interaction effect of u^{th} level of irrigation and v^{th} level of fertilizer

$e_{(uv)j}$ is the residual variation or error

Analysis

The computation of various SS is similar to that of RCBD except that the treatment SS is partitioned here as SS due main effects and interactions. In the example, the DF for treatments is 9 – 1 = 8. This can be partitioned for main effects and interaction. DF for main effect of factor I = number of levels – 1 = 3 - 1 = 2; similarly the DF for main effect of factor F = 3 - 1 = 2. DF of I x F interactions = 2 x 2 = 4. It may be verified that the sum of the DF of main effects and interaction is equal to the treatment DF.

Step 1

Compute the SS as in the case of RCBD, regardless of the factorial arrangement of the treatments.

Correction factor = $(60 + 55 + \ldots + 63 + 59)^2/27 = 99250.7$

Total SS = $60^2 + 55^2 + \ldots + 63^2 + 59^2 - 99250.7 = 996.2963$

Total of Block 1 = $60 + 55 + \ldots + 65 + 64 = 561$

Total of Block 2 = 528

Total of Block 3 = 548

Block SS = $[(561^2 + 528^2 + 548^2)/9] - 99250.7 = 61.40741$

To obtain the treatment SS, prepare the following Table 9.1.

Table 9.1. Average coconut yield for different treatment combinations

Irrigation	Fertilizer			Total
	F_0	F_1	F_2	
I_0	150	177	183	510
I_1	185	191	199	575
I_2	168	181	203	552
Total	503	549	585	1637

Treatment SS = $[(150^2 + 177^2 + 183^2 + 185^2 + \ldots + 181^2 + 203^2)/3] - 99250.7$
= $99939.67 - 99250.7 = 688.963$

Error SS = $996.2963 - 61.40741 - 688.963 = 245.9259$

Step 2

Obtain the SS due to main effects and interaction effects as below:

SS due to main effects of factor I = $(510^2 + 575^2 + 552^2)/(3 \times 3) - 99250.7$
= $99492.11 - 99250.7 = 241.4074$

SS due to main effects of factor F = $(503^2 + 549^2 + 585^2)/(3 \times 3) - 99250.7$
= $99626.11 - 99250.7 = 375.4074$

SS due to I x F interaction = Treatment SS - Factor I SS - Factor F SS
= $688.963 - 241.4074 - 375.4074$
= 72.14815

The above calculated results are arranged in the ANOVA table as given in Table 9.2.

Table 9.2. ANOVA for average nut yield

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated (0.05)
Blocks	2	61.4074	30.7037		
Treatments	8	688.9630	86.1204		
Irrigation	2	241.4074	120.7037	7.85301	3.63
Fertilizer	2	375.4074	187.7037	12.21205	3.63
Irrigation x Fertilizer	4	72.1481	18.0370	1.17349	3.01
Error	16	245.9259	15.3704		
Total	26	996.2963			

Conclusion

From the ANOVA, it can be seen that the main effects of Irrigation and Fertilizer are significantly different. The interaction effect between these two factors is not significantly different. The comparison of the different levels of the factors can be made using appropriate CD.

Note: *The interaction should be tested first and if this comes out to be significant only then proceed for further analysis for individual components.*

The CD (5%) for comparing Irrigation levels is obtained as:

$$CD = t_{\alpha,N-t}\sqrt{\left(\frac{2}{r}\right)MSE}$$

$$CD = 2.12\sqrt{\left(\frac{2}{3\times 3}\right)\times 15.3704} = 3.918$$

Note: *It may be noted that there are 3 x 3 = 9 replications for (3 blocks and 3 levels of other factor) each level of Irrigation. The same CD value can be used for comparing the levels of Fertilizer factor. The number of replications for interaction effect will only be three in this experiment and need to be remembered while calculating the CD for comparison of interaction effects.*

Analysis of factorial experiments with more than two factors

With three or more factors, besides first order interaction, the second-, third- etc. order interaction also exist. The three-factor interaction SS is obtained from the SS due to the three factor combinations by subtracting from it the SS due to all the main effects and lower order interactions of all the factors involved. Similarly, the four-factor interaction SS is obtained by first obtaining the SS due to the combinations of the four factors and subtracting from it the SS due to all the main effects and lower order interactions of all the involved factors.

In the above discussion, we have assumed that the design used is a Randomized Complete Block Design with all the combinations of the levels of the factors as

treatments. In the case of confounded experiments, the analysis is modified as follows:

If certain interactions are completely confounded, the model will not include the effects due to the confounded interaction. In case of partial confounding of certain interactions, the SS due to partially confounded interaction is obtained considering only the replications in which the interaction is not confounded and calculating the SS due to the interaction from the data of these replications only. The SS due to the lower interactions and the main effects to be subtracted for this purpose are also to be calculated from these replications only.

Many times, the higher order interactions are considered as absent when the factorial experiment involves several factors. In such a situation, the model will not involve such interaction effects that are considered as absent and the SS due to such interactions gets included as part of the error SS.

Split Plot Design

Split plot designs are a special kind of layout for conducting factorial experiments. The levels of one factor use the lay out of a standard design (e.g. RCBD) and those plots are referred as 'whole plots'. Each whole plot is further divided into small units (sub-plots) and the levels of other factor (or combination of levels of more than one factor) are allocated. Thus, each whole plot becomes a 'block' (or replication) for the subplot treatments.

When a factor requires larger plots to make it convenient for the organization of a factorial experiment, split plot design is recommended. For example, it is desired to have larger plot size for factors such as tractor operation or irrigation than that for the other factors like manure. In planning a trial with such factors, it is to be kept in mind that the precision of the sub-plot comparisons is better compared than the main plot. The randomization is restricted within blocks in such a manner that several levels of the second factor are assigned to contiguous plots with a common level of the first factor instead of scattering them over the entire block.

Example

To study the optimum fertilizer and water requirement for increased copra production in palms, a factorial experiment was conducted using split plot design. The three fertilizer levels are half, full, and one and a half quantities of recommended fertilizers for coconut and are denoted as: F_0, F_1 and F_2, respectively. The irrigation interval is the second factor and are denoted as: I_0: irrigation applied once in nine days, I_1: irrigation applied once in six days and I_2: irrigation applied once in three days.

Layout

For convenience of irrigation and also to reduce the border effects, larger plots were suggested for the irrigation treatments (main treatments). The irrigation

treatments were laid out in a RCBD with three replications. Randomization is performed as in the case of RCBD. Next, each of the whole plots is divided into three sub-plots and the fertilizer levels are allotted at random (sub-treatments). The layout of the design is given below (Fig. 9.2).

Once in six days (I_1)	Once in three days (I_2)	Once in nine days (I_0)
F_0 F_1 F_2	F_1 F_2 F_0	F_2 F_0 F_1
Once in nine days (I_2)	Once in three days (I_2)	Once in six days (I_1)
F_0 F_1 F_2	F_2 F_0 F_1	F_1 F_0 F_2
Once in three days (I_2)	Once in six days (I_1)	Once in nine days (I_0)
F_2 F_0 F_1	F_1 F_2 F_0	F_1 F_0 F_2

Figure 9.2. Field layout for split-plot experiment.

The average copra yield per palm per year calculated for each sub-plot is shown in Table 9.3.

Table 9.3. Average copra yield (kg/palm/year)

	Block-1			Block-2			Block-3		
	I_0	I_1	I_2	I_0	I_1	I_2	I_0	I_1	I_2
F_0	15.9	14.8	8.1	15.2	14.0	7.2	13.8	15.0	9.4
F_1	21.1	19.3	15.1	20.0	18.6	12.7	19.2	18.2	14.4
F_2	18.0	17.3	15.8	19.7	15.8	12.3	17.1	18.5	16.0

Model

The model for split plot design is:

$$y_{uvj} = \mu + \rho_j + i_u + e_{1uj} + f_v + (if)_{uv} + e_{2uvj}$$

where,

Y_{uvj} is the observation of v^{th} sub-plot treatment in u^{th} main plot treatment of j^{th} block;

μ is the general mean

ρ_j denotes the j^{th} block effect

i_u is the main effect of u^{th} level of irrigation

f_v is the main effect of v^{th} level of fertilizer

$(if)_{uv}$ is the interaction effect of u^{th} level of irrigation and v^{th} level of fertilizer

e_{1uj} and e_{2uvj} are the error terms, error 1 and error 2, associated with the main plots and the sub-plots, respectively.

In case of a factorial experiment, there will be only one error term in the model. But in the split plot design, there are two error terms. This is because the responses due to the sub-plot and the main-plot treatments are based on different number of replications and thus have different variances.

Analysis

The calculations of DF, SS and MS for the different sources of variation are described and illustrated below:

Degrees of freedom (DF)

The number of replications (blocks) is denoted by 'r', number of main plot treatments by 'a' and number of sub-plot treatments by 'b'.

DF for Total = N-1,

Where, N = abr = Total number of observations in an experiment

DF for main plot treatments = a-1

DF for blocks = r-1

DF for error-1 = (a-1) x (r-1)

DF for sub-plot treatments = b-1

DF for interaction (main plot x sub-plot) = (a-1) x (b-1)

DF for error-2 = (b-1) x (r-1)

In the example, a = b = r = 3.

The ANOVA of split plot design has two parts. The first part tests the significance of main-plot treatments, whereas second part tests the significance of sub-plot treatments and main-plot x sub-plot interaction. The sources of variation in the first part are obtained from the ANOVA of the design (in this example, RCBD) used for the main-plot treatments. The corresponding error SS is referred as error-1 and corresponding mean SS is used as denominator for testing the significance of the main-plot effects.

The Total SS and SS due to sub-plot treatments and the interaction between the main-plot and sub-plot treatments are obtained as in the case of a factorial experiment. Subtracting all the SS (including main-plot treatments, replications, and error-1) from the total SS will provide the error-2 SS. Corresponding MS is used as denominator to obtain the F-statistic for testing the significance of sub-plot treatments and main-plot x sub-plot treatment interaction.

Step 1

Calculation of the Correction Factor and Total SS.

Correction factor = $(15.9 + 14.8 + \ldots + 18.5 + 16)^2/27 = 6611.343$

Total SS = $15.9^2 + 14.8^2 + \ldots + 18.5^2 + 16^2 - 6611.343 = 321.6074$

Step 2

Complete the main-plot analysis as organized in Table 9.4a.

Table 9.4a. Tabulated data summary for main plot analysis

	Block 1	Block 2	Block 3	Total
I_0	55.0	54.9	50.1	160.0
I_1	51.4	48.4	51.7	151.5
I_2	39.0	32.2	39.8	111.0
Total	145.4	135.5	141.6	422.5

Block SS $= (145.4^2 + 135.5^2 + 141.6^2)/9 - 6611.343$
$= 6616.886 - 6611.343 = 5.542963$

Irrigation SS $= (160^2 + 151.5^2 + 111^2)/9 - 6611.343$
$= 6763.694 - 6611.343 = 152.3519$

Error-1 SS = Main treatment x block SS
$= (55^2 + 54.9^2 + \ldots + 32.2^2 + 39.8^2)/3 - 6611.343 -$ Block SS - Irrigation SS
$= (6782.77 - 6611.343) - 5.542963 - 152.3519$
$= 13.53259$

Step 3
Find the SS due to sub-plot treatment and interaction, using Table 9.4b below:

Table 9.4b. Tabulated data summary for sub-plot analysis

	I_0	I_1	I_2	Total
F_0	44.9	43.8	24.7	113.4
F_1	60.3	56.1	42.2	158.6
F_2	54.8	51.6	44.1	150.5

SS due to Fertilizer $= (113.4^2 + 158.6^2 + 150.5^2)/9 - 6611.343$
$= 6740.419 - 6611.343 = 129.0763$

Fertilizer x Irrigation SS $= (44.9^2 + 43.8^2 + \ldots + 44.1^2)/3 - 6611.343 -$ Fertilizer SS – Irrigation SS
$= 6906.363 - 6611.343 - 129.0763 - 152.3519 = 13.59259$

Error-2 SS = Total SS - Block SS - Irrigation SS - Error - 1 SS – Fertilizer SS – Fertilizer x Irrigation SS
$= 321.6074 - 5.542963 - 152.3519 - 129.0763 - 13.59259$
$= 7.511111$

The next step is to prepare the ANOVA as given in Table 9.5. From the ANOVA, it can be seen that main effects of irrigation and fertilizer levels as well as the interaction effects are significantly different at 5% level.

The treatment means to be compared can be arranged in the form of a table,

in which the colums represent the main plot treatments, and rows represent the sub-plot treatments as presented in Table 9.6. In split-plot design, sub-plot treatments within any given level of main plot treatment can also be made. Thus, there are four types of comparisons among treatment effects. The expression for calculation of CDs is illustrated below:

Table 9.5. ANOVA for copra yield (kg/palm/year)

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated (0.05)
Blocks	2	5.542963	2.771482		
Irrigation (Main plot)	2	152.3519	76.17595	22.5163	6.94
Error-1	4	13.53259	3.383148		
Fertilizer (Sub-plot)	2	129.0763	64.53815	103.1083	4.75
Irrigation x Fertilizer	4	13.59259	3.398148	5.428994	3.26
Error-2	12	7.51111	0.625926		
Total	26	321.6074			

(a) Comparison between main plot treatments.

$$CD_m = t_1\sqrt{\frac{2s_1^2}{br}}$$

Where, t_1 is the table value of t distribution at 5% significant level corresponding to the DF of error-1 sum of squares (s_1^2)

With regard to the above example, there will be three average values corresponding to the three levels of irrigation and the CD for comparison is,

$$CD_m = 2.776\sqrt{\frac{2\times 3.383148}{9}} = 2.406987$$

(b) Comparison between sub-plot treatments (Fertilizer levels)

$$CD_s = t_2\sqrt{\frac{2s_2^2}{ar}}$$

Where, t_2 is the table value of t distribution at 5% significant level corresponding to the DF of error-2 sum of squares (s_2^2)

$$CD_s = 2.179\sqrt{\frac{2\times 0.625926}{9}} = 0.812667$$

(c) Comparison of interaction effects (i.e. main plot means at same or different levels of sub-plots)

CD is computed as follows:

First obtain the weighted t-value as:

$$t = \frac{(b-1)t_2 s_2^2 + t_1 s_1^2}{(b-1)s_2^2 + s_1^2} = \frac{2\times 2.179\times 0.625926 + 2.776\times 3.83148}{2\times 0.625926 + 3.383148} = 2.628979$$

Now CD for comparison of interaction effects is

$$CD_{ms} = t\sqrt{\frac{2[(b-1)s_2^2 + s_1^2]}{br}} = 2.628979\sqrt{\frac{2\times(2\times 0.625926 + 3.383148)}{9}} = 2.654$$

(d) Comparison of sub-plot treatment means (Table 9.6) within a given level of main plot

$$CD_{s(m)} = t_2\sqrt{\frac{2s_2^2}{r}} = 2.179\sqrt{\frac{2\times 0.625926}{3}} = 1.40758$$

For this comparison, the average values of sub-plot treatments need to be prepared separately for the three levels of irrigation, but not attempted here.

Table 9.6. Treatment means for irrigation and fertilizer experiment in split-plot

Sub-plot (Fertilizer)	Main plot (Irrigation)			Sub-plot means
	I_0	I_1	I_2	
F_0	14.97	14.60	8.23	12.60
F_1	20.10	18.70	14.07	17.62
F_2	18.27	17.20	14.70	16.72
Main plot means	17.78	16.83	12.33	

Conclusion

From Table 9.6, it can be seen that the copra yield of I_2 is significantly less than the other two irrigation levels. The overall effects of fertilizer treatments are significant and suggest that increase or decrease of fertilizer dose other than the recommended dose will reduce the yield. The maximum yield is obtained from the combination I_0F_1 (i.e. once in three days irrigation and recommended dose of fertilizer), which is significantly better than any other treatment combinations.

Strip-Plot Design

Strip-plot design is also a kind of layout for conducting factorial experiments. In strip-plot design, the levels of one factor are superimposed over the levels of other factor at right angles. This is done first by laying out the trial with treatments constituting the levels of first factor. Next divide the blocks again into plots which are perpendicular to the plots made previously for the first factor. The plots made

later are referred as strip-plots. The levels of second factor are now assigned to the strip-plots in each block to complete the layout of the strip-plot design.

Strip-plot design is convenient in cultural management experiment involving, for instance, factors like spacing and ploughing, where the use of small plots by splitting larger plots is not feasible. A block may be divided into strips in one direction to be allotted to one set of treatments, and into another set of strips, in a direction at right angles to the first, to be allotted to the second set of treatments. In such a trial, the comparison between the levels of each of the factors allotted to the strips in the two different directions will be less precise compared to interaction between the factors.

Example

Consider an experiment to find the fertilizer requirements for three coconut cultivar/ hybrids viz., WCT, WCT x COD and COD x WCT, represented by the symbols C_1, C_2 and C_3. Three levels of fertilizers were tested (F_1, F_2, F_3). Thus, there are nine treatment combinations. It is also necessary to determine the fertilizer requirement under two soil watering schemes viz., irrigated and rainfed. The total number of treatment combinations in the experiment is therefore 18. Two limitations arise when RCBD is used. First, the block size is large and second the random arrangements of treatments in a block will result to scattering of irrigated plots thereby making it difficult to carry out the trial. Also it requires large number of border rows. An alternative is to choose a RCBD for the nine treatment combinations arising from the factorial arrangement of fertilizer and cultivars. Then consider the factor Irrigation as a strip factor and make the required number of strips across each replication (two strips in this example). The levels of the strip factor can then be randomly assigned independently in each replication.

Layout

First, the layout of the nine treatment combinations in RCBD is made. Six palms constituted a plot. Next, divide each block perpendicular to the direction by which plots were formed and randomly allocate the levels of the strip factor. In Fig. 9.3, the shaded area represents the irrigated portion of the replication while the rainfed area is the unshaded portion. The average number of nuts per plot (average of 3 palms) is shown in Table 9.7.

Model

For the reasons similar to those in the case of split-plot design, the model for strip-plot design includes three error terms. These are e_{1uj} for the strips to which the first set of treatments are allotted, e_{2vj} for the strips at right angles to the first ones and e_{3uvj} for the sub-plots formed at the cross sections of the strips in different directions as indicated below:

$$y_{uvj} = \mu + \alpha_u + \rho_j + e_{1uj} + \beta_v + e_{2vj} + (\alpha\beta)_{uv} + e_{3uvj}$$

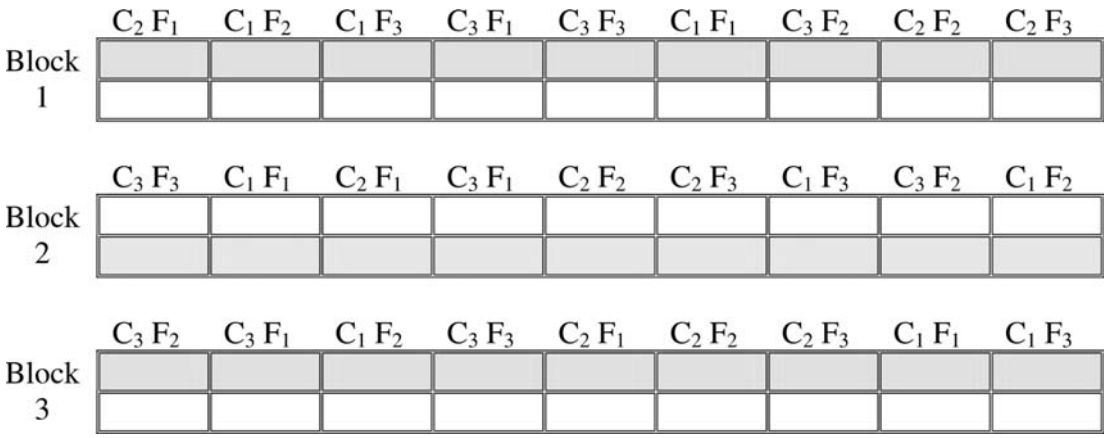


Figure 9.3. Field layout for strip-plot experiment.

Table 9.7. Data tabulated from the experiment as shown in Fig. 9.3

Fertilizer	Cultivar/ Hybrids	Rainfed/ Irrigated	Block 1	Block 2	Block 3
F_1	C_1	R	77	63	132
F_1	C_1	I	91	114	128
F_1	C_2	R	44	92	92
F_1	C_2	I	115	109	122
F_1	C_3	R	64	61	105
F_1	C_3	I	76	113	141
F_2	C_1	R	111	98	118
F_2	C_1	I	115	153	119
F_2	C_2	R	139	113	167
F_2	C_2	I	171	123	151
F_2	C_3	R	100	118	138
F_2	C_3	I	116	141	134
F_3	C_1	R	131	123	133
F_3	C_1	I	133	132	162
F_3	C_2	R	178	114	144
F_3	C_2	I	154	171	167
F_3	C_3	R	147	101	93
F_3	C_3	I	133	177	145

The effects α_u denote the effect of the u^{th} level of first set of treatments and the effects β_v is the effect of the v^{th} level of second set of treatments. In the example, the first set of treatments corresponds to the nine treatment combinations of fertilizer and cultivars. The SS due to error-1, associated with the first set of treatments is obtained as that of interaction between the replications and the first set of treatments. The SS due to error-2 is the interaction between replications and the second set of treatments. The SS due to first-, second- sets of treatments and due to their interaction are obtained as in the case of split-plot design. The SS due to the error-3 is obtained by subtracting all the other SS from the total SS. The mean sums of squares due to the errors 1, 2 and 3 are used as denominator

for the testing of the main effects of first set of treatments, main effects of the second set of treatments and their interaction, respectively.

Analysis

The calculations of DF, SS and MS for the different sources of variation are described and illustrated below:

Degrees of freedom (DF)

Denote the number of replications (blocks) by 'r', number of first set of treatments by 'a' and number of second set of treatments by 'b'.

DF for Total = N-1

Where, N = abr = Total number of observations in an experiment

DF for first set of treatments = a-1

DF for blocks = r-1

DF for error-1 = (a-1) x (r-1)

DF for second set of treatments = b-1

DF for error-2 = (b-1) x (r-1)

DF for intrection between first- and second-set of treatments = (a-1) x (b-1)

DF for error-3 is obtained by subtraction.

In the above example, a = 9; b = 2; r = 3. The computations of SS are done as follows:

The Total SS is obtained by squaring and adding the 54 observations and then subtracting the correction factor; Total SS = 50104.375.

From the Block totals (each total is based on 18 observations), obtain the Block SS.
Block SS (with 2 DF) = 3031.1527

Next, prepare the two-way table constituted by blocks as columns and the nine treatment combinations of fertilizer and cultivars as rows. From Table 9.7, the SS due to the treatments (with 8 DF) and block-by-treatment interaction can be obtained. This interaction SS is the error-1 and has 16 DF (2 x 8 = 16).

Treatment combination SS = 22977.375

Error-1 SS = 7912.8472

Following these estimates, prepare a two-way table for blocks vs. levels of strip factor (irrigation and rainfed). Note that values in the cells of this table are based on nine observations each. From this Table, obtain the SS due to strip factor (DF = 1) and also the block-by-strip factor interaction SS. This interaction SS is error 2 and has DF = 1 x 2 = 2.

SS due to irrigation (1 DF) = 6890.7453

Error-2 SS = 1824.6991

The interaction SS due to fertilizer-cultivar combination with irrigation is obtained from the respective Table and have DF = 1 x 8 = 8. Note that the value of each cell in this Table is based on three observations.

Treatment combination by irrigation SS = 1588.9212

The Error-3 is obtained by subtracting from the Total SS all other SS mentioned above and has DF = 16.

Error-3 SS = 5878.6342.

Now, the ANOVA can be prepared as shown in Table 9.8.

Table 9.8. ANOVA for number of nuts per palm

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	F-Tabulated (0.05)
Blocks	2	3031.1527	1515.5763		
Treatment combinations	8	22977.3750	2872.1718	5.81	4.46
Error-1	16	7912.8472	494.5529		
Irrigation (Strip-plot)	1	6890.7453	6890.7453	7.55	18.50
Error-2	2	1824.6991	912.3495		
Treatment-by-Irrigation	8	1588.9213	198.6151	0.54	2.59
Error-3	16	5878.6342	367.4146		
Total	53	50104.3750			

Conclusion

Based on the error term used for testing the effects in the ANOVA, the CD can be calculated. CD for comparing the treatment combinations is obtained as 27.218. Other effects are not significant in this experiment. This experiment has many shortcomings which are reflected in the ANOVA. First, the error DF for testing the factor Irrigation is only two against the requirement of 12 or more. It may be seen from the ANOVA table that despite an F value of 7.55, the irrigation factor is not significant. Also note that the plot size of three palms is too little to account for the inherent variation among palms.

References

Cochran, W.G. and Cox, G.M. 1957. Experimental Designs, 2nd edition. John Wiley and Sons, New York. 611p.

Das, M.N. and Giri, N.C. 1986. Design and Analysis of Experiments 2nd edition. Wiley Eastern, New Delhi. 488p.

Further Reading

Federer, W.T. 1955. Experimental Designs. Macmillan, New York.

Finney, D.J. 1955. Experimental Design and its Statistical Basis. Cambridge University Press, London.

Kempthorne, O. 1952. Design and Analysis of Experiments. John Wiley and Sons, New York.

Chapter 10: Analysis of multilocation trials

The performance of a variety at different locations is not the same. It can also be seen that certain varieties yield satisfactorily even in adverse conditions, while majority failed to perform. Similarly, even under good management conditions, certain varieties may not perform to the expected level. This is because, the performance or yield of a variety is not solely determined by its genetic makeup and environment alone, but modified by the genotype x environment interaction. There are several forms which genotype x environment interaction may take. Varieties responding to better management or environment are not calling for much attention. But a variety that withstanding adverse environment in which other varieties performed very badly, need to be observed closely as it could assure production under such conditions. Understanding the implication of genotype-by-environment (GE) interaction structure is therefore an important consideration in plant breeding programs.

A systematic method of generating data to meet this requirement is to conduct multi-location trials in which a set of varieties (or cultivars or genotypes) are evaluated using a standard experimental designs (often a RCBD) at various locations. Each individual experiment is analyzed separately before the data are pooled. By pooling the data from all the trials, it is possible to generate information on the adaptation and yield stability of the cultivars. The methods used in this regard are discussed in this chapter.

Only a few multi-location trials in coconut experimentation are being carried out so far. In India, as part of the All India Coordinated Research Project on Palms, hybrid/cultivar evaluation trials are laid out at different agroclimatic regions of the country. A regional example is the COGENT trial involving six coconut hybrids being tested in seven countries (Brazil, Jamaica, Mexico, Cote d'Ivoire, Benin, Tanzania and Mozambique). In these trials, RCBD is being used.

Genotype and environment

With regard to comparison of varieties in a set of multi-location trials, the term 'genotype' refers to a cultivar and the term 'environment' relates to the set of climatic, soil, biotic, and management conditions in a trial conducted at a particular location (Annicchiarico 2002). In the absence of genotype x environment interaction, a comparison of genotypes on the basis of average performance over different environments will indicate the better performing genotype(s). However, if genotype x environment interaction is significant, it is indicative that differences between genotypes vary widely across the environments. To study the differential performance of a set of genotypes tested at different locations, we use the concept of stability.

Concept of stability

There are at least three different ways that stability can be defined (Lin *et al*. 1986), among which, the type 2 or agronomic stability is of relevance in making recommendations on a genotype. A genotype is considered to be stable if its response to environments is parallel to the mean response of all genotypes in the trial. This is also referred as the dynamic concept of stability (Becker and Leon 1988). The deviation of a genotype from the average performance of other genotypes is considered here as a contribution to 'instability'. Stability defined in this way is clearly a measure, relative to the genotypes included in the test so its scope of inferences is confined to the test genotypes. In other words, a genotype considered to be stable across environments by this definition is true only with respect to the other tested genotypes.

Note: *Instability need not be a bad thing. If the breeders are looking for niche-specific genotypes/cultivars, one that may be unstable, but performs well consistently over the years can be recommended to be grown in that particular environment as against more stable genotypes as stability may not be correlated with high yield.*

Model

Let X_{ijk} denote the yield in the k^{th} replicate of the i^{th} genotype in the j^{th} location (environment). We assume that the design is RCBD with 'r' replications and there are 'g' genotypes tested in 's' locations. The model will be:

$$X_{ijk} = \mu + \tau_i + \varepsilon_j + \gamma_{ij} + \rho_{k(j)} + e_{ijk}$$

Where,

μ is the general mean
τ_i is the effect of genotype i
ε_j is the effect of environment j
γ_{ij} is the effect of genotype x environment interaction
$\rho_{k(j)}$ is the k^{th} block effect within location j
e_{ijk} is the residual variation or error assumed to be normally distributed with mean 0 and variance σ^2, [i.e. $e_{ijk} \sim N(0, \sigma^2)$].

Testing of effects

The different sources of variation are obtained from the analysis of variance (ANOVA) as shown in Table 10.1 (adopted from Annicchiarico, 2002).

The sum of squares due to blocks within a location and pooled error are obtained by summing the relevant terms from individual analysis of variance carried out for each location. ANOVA for RCBD is described in Chapter 7. To obtain the sum of squares due to genotype, location and genotype x location, Table 10.2 may be used. Each cell of the Table 10.2 is the sum of *r* observations (i.e., X_{ij}). The marginal totals $X_{i.}$ is the sum of *s.r* observations and $X_{.j}$ is the sum of *g.r* observations.

Table 10.1. Analysis of variance of multi-location trial

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value
Genotype	g-1	*(Computed from two-way*	M_g	M_g / M_e
Location	s-1	*table of genotype x*	M_l	M_l / M_e
Genotype x Location	(g-1) (s-1)	*location)*	M_{gl}	M_{gl} / M_e
Blocks within location	s (r-1)	*(Summing over individual RCBD analysis at each*	M_b	
Error	s (g-1)(r-1)	*location)*	M_e	

Table 10.2. Data summary for genotypes and locations

Genotype	Location			Total
	1	....	s	
1	X_{11}	....	X_{1s}	$X_{1.}$
...	....	....	....	....
g	X_{g1}	....	X_{gs}	$X_{g.}$
Total	$X_{.1}$	....	$X_{.s}$	X..

SS due to genotype is GSS $= (X_{1.}^2 + .. + X_{g.}^2)/(s.r) - X..^2/(g.s.r)$

SS due to locations is LSS $= (X_{.1}^2 + .. + X_{.s}^2)/(g.r) - X..^2/(g.s.r)$

SS due genotype x location interaction $= (X_{11}^2 + .. + X_{gs}^2)/r - X..^2/(g.s.r) - GSS - LSS$

Measures of stability

Once the genotype x location interaction is found to be significant, the most adaptable or stable genotype needs to be identified. The most commonly followed procedure is the joint linear regression modelling and complementary analysis (Annicchiarico 2002). In this approach, the genotypes x environment interaction effects are expressed in the form of a regression equation with the environmental effect. There are two ways to form the regression equation:

(1) $\gamma_{ij} = \beta_i \varepsilon_j + d_{ij}$ (Perkins and Jinks 1968)

Where,

β_i is the regression coefficient of the genotype i on environment

ε_j is the effect of environment j, and

d_{ij} is the deviation from the regression.

(2) $m_{ij} = a_i + b_i m_j + d'_{ij}$ (Finlay and Wilkinson 1963)

Where,

m_{ij} is the average value of i^{th} genotype in the j^{th} environment (can be obtained from Table 10.2 as $m_{ij} = X_{ij}/k$)

a_i is the constant term in the regression equation
b_i is the regression coefficient of the genotype i on average values of environments
m_j is the average value for j^{th} environment [$(m_j = X_{.j}/(g.r)$], and
d'_{ij} is the deviation

The regression coefficients in the above equations are equivalent as: $\beta_i = b_i - 1$. The expression for b_i is given below:

$$b_i = 1 + \sum_j (X_{ij} - m_i - m_j + m)(m_j - m)/\sum_j (m_j - m)^2$$

where, $m_i = X_{i.}/(s.r)$ and $m = X../(g.s.r)$; and summation is over all environments.

The average regression coefficients for all genotypes (β_i s) will be zero. Obviously, the average value of b_i is equal to 1. Large positive β values, if associated with relatively high mean yield, will result in specific adaptation to high-yielding locations. On the other hand, large negative β values associated with relatively high mean yield result in specific adaptation to low-yielding (unfavourable) sites. A value of β around zero indicates a lack of specific adaptation. A genotype with zero β and high yield is considered suitable for all locations.

Consequent to the introduction of regression terms in the model, the SS due to genotype x location as mentioned in the ANOVA (Table 10.1) can be partitioned further as:

A. The location SS partitioned into:
   1. Combined regression SS with 1 DF (by fitting a common regression of the response variable (e.g. yield) on the environment index) with 1 DF; and
   2. Residual SS (by subtracting the combined regression SS from the location SS).

B. The genotype x location SS partitioned into:
   1. Heterogeneity between regressions with (*g-1*) DF [by subtracting the combined regression SS from the sum of individual regression SS (based on separate regression analysis for each genotype)]; and
   2. Residual SS (by subtracting the heterogeneity SS from genotype x location SS) with (*g-1*)(*s-2*) DF.

The environment index is usually taken as the deviation of the location mean from the overall mean (i.e. $m_j - m$, and is denoted by z_j). The computation of SS is shown in Table 10.3.

Illustration

Consider a multilocation trial on coconut with six cultivars grown in four locations. The design used was RCBD with four replications. The yield data is shown in Table 10.4.

Table 10.3. Partitioning of sum of squares according to combined regression analysis

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-value
Genotype	g-1	As in Table 11.2	M_g	M_g/M_e
Location				
Combined regression	1	$(\Sigma_i \Sigma_j\ m_{ij} z_j)^2/\Sigma_i\ \Sigma_j\ z_j^2$	M_{l1}	M_{l1}/M_e
Residual	s-2	By subtracting the above SS from location SS	M_{l2}	M_{l2}/M_e
Genotype x Location				
Heterogeneity of regressions	g-1	$\Sigma_i\ [(\Sigma_j\ m_{ij} z_j)^2/\Sigma_j\ z_j^2]$ – combined regression SS	M_{gl1}	M_{gl1}/M_e
Reminder	(g-1)(s-2)	By subtracting the above SS from G x L SS	M_{gl2}	M_{gl2}/M_e
Blocks within location	s (r-1)		M_b	
Error	s (g-1)(r-1)		M_e	

Table 10.4. Yield data from a multi-location trial of coconut (nuts/palm/year)

Location	Cultivar	Replications			
		1	2	3	4
Location-1	LCT	40.60	77.00	24.50	52.50
	CCT	58.40	75.70	58.00	86.50
	AOT	78.00	91.80	81.40	72.00
	PHOT	90.00	105.00	86.90	99.50
	FJT	92.00	123.90	61.10	97.20
	JMT	25.90	50.40	46.90	22.00
Location-2	LCT	97.90	114.00	94.40	74.40
	CCT	76.00	75.00	43.20	58.70
	AOT	48.00	88.30	79.80	69.20
	PHOT	94.00	55.60	87.30	77.90
	FJT	80.90	77.90	88.00	73.00
	JMT	44.70	28.30	47.00	39.20
Location-3	LCT	107.00	132.00	117.00	115.30
	CCT	122.00	129.00	111.00	110.00
	AOT	94.00	187.00	153.00	138.70
	PHOT	96.70	132.00	105.00	121.20
	FJT	98.20	106.00	115.00	95.80
	JMT	73.00	97.00	124.00	118.80
Location-4	LCT	97.80	97.00	96.90	93.90
	CCT	114.00	93.20	102.80	80.60
	AOT	129.30	132.70	143.70	90.00
	PHOT	100.70	81.20	123.50	80.20
	FJT	82.70	64.00	76.20	84.40
	JMT	102.60	110.30	108.80	92.10

The ANOVA as indicated in Table 10.1 was worked out and presented in Table 10.5. It can be seen that genotype x location effect was significant. Hence, the cultivars were compared for stability for which we attempted the regression analysis.

In the regression analysis, we attempted the prediction of response of a genotype in a location based on the environmental index of that location (which could be

an independent measure or derived from the data as the deviation of the location mean from the overall mean (i.e. m_j - m). In this example, we used the later as the environmental index. The regression analysis was done in two stages: For each genotype regression coefficients were obtained separately (i.e. 6 regression analysis in this example). The regression coefficients, regression SS and residual SS of 'individual' regression analysis for the six cultivars are shown in Table 10.6.

Table 10.5. Analysis of variance for coconut yield

Sources of variation	DF	Sum of squares	Mean sum of squares	F-value	F-Tabulated (0.05)**
Genotype	5	10201.14	2040.227	9.8182**	3.34
Location	3	36220.33	12073.44	58.1011**	4.13
Genotype x Location	15	17279.46	1151.964	5.5436**	2.35
Blocks within Location	12	7814.594	651.2162	3.1338	2.50
Error	60	12468.05	207.8008		

**significant at 1%.

Table 10.6. Estimates of regression coefficients and corresponding regression sum of squares

Estimates	Genotypes						
	AOT	CCT	FJT	JMT	LCT	PHOT	Total
Regression coefficient	1.51*	1.12**	0.24	1.59	0.99	0.54	6.00
Regression SS	3451.00*	1914.80**	86.59	3824.34	1488.15	435.17	11200.05
Residual SS	83.94	25.78	380.87	442.63	1062.69	178.86	2174.77

*Significant at 5%; **Significant at 1%

Irrespective of cultivars, a single regression was obtained (i.e. the combined regression coefficient, equal to 1 when environmental index is defined as a 'deviation from mean' as in this example). The combined regression coefficient was obtained as 1 and the corresponding SS as 9054.67 (with 1DF). The ANOVA is presented in Table 10.7.

From Table 10.6, it can be seen that the regression coefficients of AOT and CCT are significantly different and also the respective regression SS. The combined regression SS and its deviation from the location SS (i.e. Residual SS as shown in Table 10.7) are significant. Therefore, not all the cultivars are responding in similar manner across environments. More specifically, a significant portion of cultivar variability across locations is not predictable.

With regard to partitioning of the location x genotype interaction SS to heterogeneity between regression and remainder, only the latter SS is found to be significant. Thus, it further confirmed the non existence of simple linear relationship between the genotype responses in an environment with respective index. Therefore, no prediction of cultivar performance could be made by this approach. [Prediction of cultivar performance is possible in a situation where the 'heterogeneity SS' alone is significant. If both the SS are significant, the practical usefulness of any prediction will depend on the relative magnitudes of the two MSs (Perkins and Jinks 1968)].

Table 10.7. ANOVA for regression analysis

Sources of variation	DF	Sum of squares	Mean sum of squares	F-value
Genotype	5	10201.14	2040.23	9.82**
Location	3	36220.33	12073.44	58.101**
Combined regression	1	9054.67	9054.67	43.57**
Residual	2	36220.33 - 9054.67 = 27165.66	13582.83	65.36**
Genotype x Location	15	17279.46	1151.96	5.54**
Heterogeneity of regressions	5	11200.05 - 9054.67 = 2145.38	429.076	2.06
Reminder	10	17279.46 - 2145.48 = 15133.98	1513.40	7.28**
Blocks within location	12	7814.59	651.22	3.13**
Error	60	12468.05	207.80	

**Significant at 1%

$F_{(0.01)}$ tabulated values for DF (1,60); (2,60); (3,60); (5,60); (10,60); (12,60); and (15,60) are 7.08; 4.98; 4.13; 3.34; 2.63; 2.5; and 2.35, respectively. $F_{(0.05)}$ tabulated value for DF (5,60) is 2.37.

Even if the heterogeneity SS is not significant, there could be few genotypes having linear association with environmental values and reliable predictions can be made for such varieties. Lin *et al*. (1986) suggested that the estimated variance of genotype deviations from the regressions (s_d^2) [the 'second stability measure of Eberhart and Russell (1966)] may be used as an indicator of goodness of fit of the (individual) regression model; a large s_d^2 value indicates a poor fit. In such situations, other measures of stability may be considered and the simplest being Wricke's (Wricke 1962) ecovalence (W_i^2) which is defined as:

$$W_i^2 = \sum_{j=1}^{s} (m_{ij} - m_i - m_j + m)^2$$

Cultivars with low values of W_i^2 are considered as more stable ones. The aforesaid stability measures are shown in Table 10.8 along with mean yield of cultivars.

Table 10.8. Estimates of stability parameters

Genotype	AOT	CCT	FJT	JMT	LCT	PHOT
b_i (Finlay and Wilkinson 1963)	1.51	1.13	0.24	1.59	0.99	0.54
b_i (Perkins and Jinks 1968)	0.51	0.13	-0.76	0.59	-0.01	-0.46
s_d^2 (Eberhart and Russell 1966)	41.97	12.89	190.43	221.31	531.34	89.43
W_i^2 (Wricke 1962)	479.90	49.90	1253.60	971.30	1062.80	502.40
Mean yield (nuts/palm/year)	104.80	87.10	88.50	70.70	89.50	96.00

With regard to regression coefficient across environment (b_i), genotypes are considered to be stable for values close to one. In other words, when the absolute difference of b_i from unity (i.e. $|b_i - 1|$) increases, the genotype is considered to be unstable. This is equivalent to saying that cultivars with value of β_i near

to zero are stable. Accordingly, from Table 10.8, it can be seen that the cultivars LCT and CCT are more stable. However, the utility of regression coefficient as a measure of stability for the cultivar LCT is questionable because of insignificant regression coefficient and regression SS (Table 10.6). This is further evident from the high value of s_d^2 (which is mean residual SS) for that cultivar. Under such circumstances, the other stability measures need to be utilized for measuring stability. The Wricke's ecovalence (W_i^2) for LCT is very high compared to that of CCT (Table 10.8). Hence, we may conclude that among the six cultivars compared, the most stable cultivar is CCT. However, the recommendation of a cultivar should not be made solely on the basis of its stability. The yield performance of the cultivars should be taken into account. As shown in Table 10.8, the cultivar CCT has relatively less yield compared with the cultivars AOT and PHOT, which are also relatively stable compared to other cultivars.

Due to its simplicity, the joint regression model has been the most popular approach for analysis of adaptation (Becker and Léon 1988). However, the method has certain limitations under situations where: (i) extreme values of site mean yield are represented by just one location; (ii) non-linear genotype responses to environment mean, etc. Hence many other approaches were proposed for genotype x environment analysis. One of the approaches is to fit a multiplicative model for interaction component (γ_{ij}), keeping the main effects additive as in the original model. This is referred as AMMI (additive main effects and multiplicative interaction) model. In AMMI analysis, first the main effects (i.e. genotype and environment) are estimated based on the ANOVA, and then the deviations (D_{ij}) are partitioned into:

$$D_{ij} = \sum \lambda_{(G)n} \upsilon_{in} \lambda_{(E)n} \xi_{jn} + e_{ij}$$

Where, υ_{in} and ξ_{jn} are eigenvectors (scaled as unit vectors, i.e. $\sum \upsilon_{in}^2 = \sum \xi_{in}^2 = 1$) of the genotype i and the environment j, respectively, and $l_{(G)n}$ and $l_{(E)n}$ are the corresponding singular values (i.e. the square root of the latent root or eigen value) for the principal component (PC) axis n; and e_{ij} is the deviation from the model. Principal component analysis is explained in the next Chapter.

There are several possible AMMI models characterized by a number of significant PC axes ranging from zero (AMMI-0, i.e. additive model) to a minimum between (g - 1) and (l - 1), where g = number of genotypes and l = number of locations. The full model (AMMI-F), with the highest number of PC axes, provides a perfect fit between expected and observed data. Models including one (AMMI-1) or two (AMMI-2) PC axes are usually the most appropriate where there is significant GL interaction. Due to their simplicity, they provide a notable reduction of dimensionality for the adaptation patterns relative to observed data. While principal components analysis is usually executed on the correlation matrix, for AMMI modelling it is executed on the covariance matrix. Furthermore, two (not one) analyses are performed simultaneously: in the analysis the genotypes are individuals (rows) and the

locations original variables (columns); in the other, vice versa (Annicchiarico 2002). Illustration of AMMI analysis is beyond the scope of this manual.

References

Annicchiarico, P. 2002. Genotype x environment interactions - challenges and opportunities for plant breeding and cultivar recommendations. FAO, Rome. (http://www.fao.org/documents/show_cdr.asp?url_file=/DOCREP/005/Y4391E/y4391e0c.htm).

Becker, H.C. and Leon, J. 1988. Stability analysis in plant breeding. Plant Breeding. 101: 1-23.

Eberhart, S.A. and Russel, W.A. 1966. Stability parameters for comparing varieties. Crop Sci. 6: 36-40.

Finlay, K.W. and Wilkinson, G.N. 1963. The analysis of adaptation in a plant-breeding programme. Aust. J. Agric. Res. 14: 742-754.

Lin, C.S., Binns, M.R. and Lefkovitch, L.P. 1986. Stability analysis: Where do we stand? Crop Science. 26: 894-900.

Perkins, J.M. and Jinks, J.L. 1968. Environment and genotype x environmental components of variability III. Multiple lines and crosses. Heredity 23: 339-356.

Wricke, G. 1962. Uber eine Methode zur Erfassung der okologischen Streubrete in Feldversuchen. Z. Pflanzenzûchtg. 47: 92-96.

Chapter 11: Multivariate analysis and determination of genetic distance

Genetic distance estimates are mostly based on morphological, biochemical (isoenzymes) and molecular markers, though more and more emphasis is on the use of molecular markers as these seem to be less affected by environment and thus the error due to G x E interactions can be minimized. Generally, analytical methods include Mahalanobis' distance, genealogical distance, use of pedigree information, multivariate distance, kinship coefficient (dos Santos Dias *et al*. 2004). All these methods can be used irrespective of markers used, though when molecular markers are used, simple clustering techniques can provide the required information on genetic distance.

In the case when specific markers that can provide information on the presence or absence of alleles are used, as in the case of most of the molecular markers (in which case data would be binary in nature), then there are four popular genetic distance measures: the D_A distance (Nei *et al*. 1983), the chord distance, D_C (Cavalli-Sforza and Edwards 1967), the standard genetic distance of Nei, D_S (Nei 1972, 1978) and the Weir and Cokerham estimator of F_{ST}, θ (Weir and Cokerham 1984). However, we will be limiting our discussion in this Chapter to the use of mostly morphological data, without much inference on the number of alleles involved. In addition, use of molecular tools is of recent origin in coconut and most of accession descriptions are morphological and quantitative in nature, D^2 statistic will be quite adequate. Readers who are interested in these genetic distance statistics are referred to the citations given in this chapter.

In the previous chapters we discussed comparison of treatments based on a single response variable. However, in many situations, the experimenter may wish to make comparisons with more than one character (not necessary a gene or allele as a single trait may be governed by more than one gene and may have several alleles). For example, in germplasm evaluation trials, we would like to evaluate accessions for different yield-related characters such as number of bunches, fruit setting, copra content, total copra yield, different fruit components, etc. Obviously, such characters are expected to have correlation between them and in such a situation multivariate analysis methods are most appropriate. This is because, a series of univariate analysis (a single variable at one time) carried out separately on each variable may lead to incorrect conclusions as the correlations or interdependence among the variables is ignored. On the other hand, the multivariate techniques are concerned with the relationships of inter-related variables.

The multivariate techniques can be broadly grouped into: Dependence methods and Interdependence methods. If variables of one set were the realization of certain dependent or criterion measures, the appropriate techniques would be the dependence methods. Important dependence methods include multiple regression, multivariate analysis of variance, discriminant analysis, canonical analysis and logit analysis.

If no distinction is made among the variables, we opt for interdependence methods for the analysis that include principal components analysis, factor analysis, multi-dimensional scaling and log linear models (Dillon and Goldstein 1984).

Among the dependence methods, the multiple regression procedure is already discussed in Chapter 5. In this chapter we will discuss the multivariate analysis of variance (MANOVA) which is an extension to ANOVA. Once the accessions (treatments) are found to be significant based on MANOVA, the interests arise on pair wise comparison between treatments and also forming of groups of treatments or accessions that are similar in nature. Mahalanobis' generalized distance is used for pair wise comparison and also used as a measure of dissimilarity for cluster analysis to determine the genetic distance among accessions as described in the subsequent sections of this chapter.

MANOVA

It is used for the simultaneous test of equality of sets of means as against individual means specified in ANOVA. The data are assumed to be drawn from a multivariate normal population with the same variance-covariance matrix in each group.

Null hypothesis

Analogous to that of ANOVA (which is used to test the hypothesis that the samples came from populations having the same mean (i.e. $\mu_1 = \mu_2 = \ldots = \mu_k$), we formulate the null hypothesis for test based on MANOVA. The mean of the characters is represented in the form of a column vector and is denoted by μ_i for the i^{th} population; same order is followed for each population. The null hypothesis is then:

H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$

Test statistic

There are many test statistics proposed for determining the equality of mean vectors of populations following multivariate normal distribution. One commonly followed test is Wilk's lambda (Λ), which is the ratio of the determinants of the matrices **W** and **W+D**; i.e. $\Lambda = |\mathbf{W}| / |\mathbf{W}+\mathbf{D}|$

Where, **W** and **D** are the sum of products matrices due to 'residual' and 'deviation from hypothesis'. The determinant of matrix is a scalar quantity and can be obtained by using the MDETERM function of MSEXCEL or specific computer programmes available elsewhere.

For example, in the case of one-way classified data, (from CRD), **D** represents the 'between treatment' sum of products matrix and **W** is the matrix of sum of products against the 'error' term. In the case of RCBD, the matrix **D** is again the 'between treatment' sum of products matrix (for testing the significance of treatments) and **W**, the corresponding error sum of products matrix (Johnson and Wichern 1992).

Based on Λ, the test statistic is developed as a function of natural logarithm of Λ as:
$- m \log_e \Lambda$.

Where, m = error DF - [(p+1) - hypothesis DF]/2;
p = number of variables.

Under null hypothesis, the above test statistic has a chi-square distribution (approximately) with DF = p times the hypothesis DF.

In the case of CRD with k treatments and a total of n experimental units,
$m = n - k - [p+1 - (k-1)]/2$
$= n - 1 - (p+k)/2$
and under the null hypothesis, the test statistic has a chi-square distribution with DF = p(k-1).

Example

Data on four fruit characters of four palms each from five coconut accessions are shown in Table 11.1. It is desired to test whether the accessions are significantly different with respect to the fruit characters. [This data is being analyzed as a one-way classified data (i.e., analogous to CRD].

Table 11.1. Fruit characters of five coconut accessions

Palms	Accessions	Fruit weight (g)	Fruit length (cm)	Husk thickness (cm)	Husk weight (g)
1	SSAT	984.50	26.875	2.250	245.8
2	SSAT	1040.00	32.500	2.725	329.0
3	SSAT	712.00	25.875	2.250	192.8
4	SSAT	1100.25	28.750	2.475	280.0
5	POLT	765.25	27.875	3.650	270.0
6	POLT	713.67	29.500	3.900	333.3
7	POLT	669.50	28.125	3.425	271.5
8	POLT	629.50	29.250	3.600	272.5
9	MVT	1591.67	33.833	3.167	373.3
10	MVT	1589.25	32.250	2.925	384.3
11	MVT	2372.50	35.250	4.225	893.5
12	MVT	1723.25	33.500	3.300	502.0
13	KKT	1407.50	32.250	2.600	401.8
14	KKT	1863.75	34.125	3.550	615.0
15	KKT	1069.50	29.375	2.875	328.3
16	KKT	1395.50	31.500	3.225	475.5
17	NLAD	980.75	30.500	3.225	413.3
18	NLAD	963.50	31.250	3.062	432.5
19	NLAD	1047.25	31.500	2.925	410.8
20	NLAD	1056.50	31.375	3.362	472.3
	Total	23675.59	615.458	62.716	7897.5

The computations will involve the following steps:

Step 1
Analogous to the 'between treatments' (accessions) sum of squares in the case of ANOVA, obtain the between sum of squares and sum of products matrix. In the matrix, the diagonal elements are the sum of squares; only the upper or lower portion of the matrix needs to be computed. For the sake of illustration, the computations of one diagonal and one off-diagonal element of the matrix are explained below.

Computation of the diagonal element corresponding to fruit weight

It is the equivalent to the various SS calculated for ANOVA for that character. For example, the correction factor is obtained as 23675.59 x 23675.59/20 = 28026678.1. The total SS, between accessions SS and error SS are then obtained as described in Chapter 7 and are 3987563.3, 3142893.0 and 844670.3, respectively. Similarly, the other diagonal elements against fruit length, etc. can be computed.

Computation of the off-diagonal elements corresponding to fruit weight and fruit length

The correction factor to be subtracted while computing the various sum of products (SP) between these two characters is: 23675.59 x 615.458/20 = 728566.6.

The Total SP = (984.50 x 26.875 + …. + 1056.50 x 31.375) - 728566.6 = 18363.5

To calculate the between accession SP of the fruit weight and fruit length, first obtain the sum of values for each accession for the characters as shown below:

Accessions:	**SSAT**	**POLT**	**MVT**	**KKT**	**NLAD**
Fruit weight	5736.25	7276.67	4048.00	2777.92	3836.75
Fruit length	127.25	134.83	124.63	114.75	114.00

The required SP is obtained as:

(5736.25 x 127.25/4 + 7276.67 x 134.83/4 + …+
3836.75 x 114.00/4) - 728566.6 = 14361.2

The error SP is then obtained by subtracting the required SP (14361.2) from the total SP (18363.5), giving the value 4002.3. In a similar manner the other elements of the matrices can be obtained as shown in Table 11.2.

Step 2
Next step is to obtain the determinant of the matrices **W** and **W**+**D** by using function MDETERM of MSEXCEL. The determinant of the error SSSP matrix (**W**) is obtained as 242973699302.8. The determinant of the matrix sum **W** + **D** (i.e. the total SSSP

matrix in this case) is obtained as 42999447542307.5. The Λ statistic is then the ratio of these two determinants and is equal to 0.00565.

Step 3

To obtain the test statistic, compute:

$m = n\text{-}1 - (p+k)/2$

$= 20 - 1 - (4+5)/2 = 14.5;$

$- m \log_e \Lambda = - 14.5 \times \log_e 0.00565$

$= - 14.5 \times -5.1761 = 75.05345$

Table 11.2 Sum of squares and sum of products (SSSP) matrices of MANOVA

Matrix of	Characters	Fruit weight	Fruit length	Husk thickness	Husk weight
Correction	Fruit weight	28026678.1	728566.6	74241.9	9348898.6
factor	Fruit length	728566.6	18939.4	1930.0	243029.0
	Husk thickness	74241.9	1930.0	196.7	24765.0
	Husk weight	9348898.6	243029.0	24765.0	3118525.3
Total SSSP	Fruit weight	3987563.3	18363.5	1247.4	1145876
	Fruit length	18363.5	121.7	10.6	6056.5
	Husk thickness	1247.4	10.6	5.3	897.6
	Husk weight	1145876.1	6056.5	897.6	458294.6
Between	Fruit weight	3142893.0	14361.2	248.0	730073.7
accessions	Fruit length	14361.2	77.4	5.1	4101.9
SSSP	Husk thickness	248.0	5.1	3.4	296.0
	Husk weight	730073.7	4101.9	296.0	219795.2
Error SSSP	Fruit weight	844670.3	4002.3	999.4	415802.4
	Fruit length	4002.3	44.3	5.5	1954.6
	Husk thickness	999.4	5.5	1.9	601.6
	Husk weight	415802.4	1954.6	601.6	238499.4

Under the null hypothesis, the test statistic has approximate chi-square distribution with

$DF = p\ (k-1)$

$= 4 \times (5-1) = 16.$

Decision: Since $\chi^2_{16,0.05} = 26.296 < 75.053$, we reject the null hypothesis.

Conclusion

The accessions are significantly different with respect to the four fruit characters.

Grouping of accessions

Once the accessions are found to be different, the next step is to find out similar groups. It is easy with regard to single character using the critical difference (CD). Such an approach will not be practical while considering many characters simultaneously because with respect to each character, different groups will be obtained making the interpretation a difficult task. A solution to this problem is

to group the accessions in terms of some similarity or dissimilarity coefficients obtained based on the values of the characters. Obviously, the characters that do not distinguish the accessions need not be considered while computing such coefficients.

A most commonly used dissimilarity coefficient is the Euclidean distance, which is the straight line distance between two points. Suppose the ordinates of the two points, A and B in a plane are denoted as (a_1, a_2) and (b_1, b_2), then the Euclidean distance is given by:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

In practice we will be having more than two characters, say *p* and the values of two individuals (accessions) are represented as (a_1, a_2, …, a_p) and (b_1, b_2, …., b_p). Then the Euclidean distance between the individuals is computed as:

$\sqrt{\sum_i (a_i - b_i)^2}$, in which the summation is for all the *p* characters.

The Euclidean distance is defined for situations where the axes are perpendicular (un-correlated). This is not the case with biological observations where characters are often correlated. In such cases, the Euclidean distances will not provide accurate spatial relationships between the accessions. It is therefore necessary to eliminate the correlation between variables before finding the Euclidean distance. The Mahalanobis' generalized distance achieves this by means of transforming the original (correlated) variable to uncorrelated variables (Rao 1952). Alternatively, new set of independent variables can be created by means of principal component analysis to allow the computation of the Euclidean distance. Obviously, the distance obtained through these two approaches will not be the same.

Mahalanobis' D^2

Assuming that the covariance matrix (**S**) is the same for all populations (accessions), the generalized distance D^2 between the two populations i and j is defined as:

$$D^2_{ij} = (\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j)$$

When the population parameters are estimated by sample statistics, the estimate of D^2_{ij} is obtained as:

$$d^2_{ij} = (\bar{x}_i - \bar{x}_j)' S^{-1}(\bar{x}_i - \bar{x}_j)$$

where,
$\bar{x}_i$ is the vector of sample mean for i^{th} population and **S** is the sample covariance matrix.

Example

With regard to data on fruit characters of five accessions (Table 11.1), we have already worked out the error SSSP matrix. Dividing each element of the matrix with error DF (15), we get the matrix **S**. Using the MINVERSE function of MSEXCEL (or any other subroutines for finding out the inverse of a matrix), $\mathbf{S}^{-1}$ can be obtained. From the example, the values for $\mathbf{S}^{-1}$ are shown in Table 11.3.

Table 11.3. Elements of the inverse of matrix S (i.e., S^{-1})

Characters	Fruit weight	Fruit length	Husk thickness	Husk weight
Fruit weight	0.00015	-0.00376	0.02557	-0.0003
Fruit length	-0.00376	0.63847	-1.43725	0.00495
Husk thickness	0.02557	-1.43725	44.66112	-0.14546
Husk weight	-0.00030	0.00495	-0.14546	0.00092

The mean values of characters for the different accessions and the difference of mean values for the accessions KKT and MVT are shown in Table 11.4.

Table 11.4. Average values for fruit characters

Accessions	Fruit weight	Fruit length	Husk thickness	Husk weight
SSAT	959.188	28.500	2.425	261.900
POLT	694.480	28.688	3.644	286.825
MVT 1819.168	33.708	3.404	538.275	
KKT 1434.063	31.813	3.063	455.150	
NLAD	1012.000	31.156	3.144	432.225
Difference of means between SSAT and POLT	264.708	-0.188	-1.219	-24.925

The next step is to work out the $d^2_{SSAT,POLT}$ as shown below:

$$(\bar{x}_i - \bar{x}_j)'\mathbf{S}^{-1} =$$

$$\begin{bmatrix}264.708 & -0.188 & -1.219 & -24.925\end{bmatrix} \begin{bmatrix}0.000154 & -0.00376 & 0.025573 & -0.0003 \\ -0.00376 & 0.638466 & -1.43725 & 0.004954 \\ 0.025573 & -1.43725 & 44.66112 & -0.14546 \\ -0.0003 & 0.004954 & -0.14546 & 0.000917\end{bmatrix}$$

$$= \begin{bmatrix}0.01538276 & 0.599829 & -49.071 & 0.090468\end{bmatrix}$$

Therefore, $d^2_{SSAT,\ POLT} = (\bar{x}_i - \bar{x}_j)'\mathbf{S}^{-1}(\bar{x}_i - \bar{x}_j)$

$$= \begin{bmatrix} 0.01538276 & 0.599829 & -49.071 & 0.090468 \end{bmatrix} \begin{bmatrix} 264.708 \\ -0.188 \\ -1.219 \\ -24.925 \end{bmatrix}$$

$$= 61.5098$$

In a similar manner, we can work out the distance between all the 5 x 4/2 = 10 pairs of accessions as shown in Table 11.5.

Table 11.5. Mahalanobis' generalized distance between the five coconut accessions

	SSAT	POLT	MVT	KKT	NLAD
SSAT	0				
POLT	61.50979	0			
MVT	31.60340	79.71975	0		
KKT	6.90998	59.40381	10.3132	0	
NLAD	13.54683	40.17097	54.6633	19.72299	0

Test of significance: With equal sample size r, the test statistic is

$$\frac{r(r-p)}{2p(r-1)} d^2 \sim F_{(p,r-p)}$$

Note: *In the above example, the number of observation per accession is four and therefore r - p = 4 - 4 = 0 and hence, the above mentioned test is not possible. To perform the test of significance of the generalized distance, the condition r – 1 ≥ p need to be hold good. In other words to test the significance of D^2, between two populations, the sample size (for both the populations) must be greater than the number of characters. In practice this condition seldom satisfies especially when there are many populations to be studied. Nevertheless, the d^2 values can be computed for situations where the condition n-k ≥ p is satisfied and can be used as a measure of dissimilarity for clustering analysis. This condition puts only a simple restriction that the total number of individuals (or units) sampled from all the populations should be greater than the sum of number of characters and number of populations. Reducing the number of populations in a study may not be practical, but the investigator can chose the number of characters for computation of d^2 to satisfy the restriction mentioned above.*

Cluster analysis

Even if statistical tests of significance are not applied, the generalized distance could be used for grouping the populations (accessions) by means of cluster analysis. When number of populations is not very large (say, less than 50), hierarchical classification is attempted and the results are summarized in the form of a dendrogram. If the number of populations is more, the construction and

representation of the dendrogram becomes tedious and in such situations non-hierarchical cluster analysis may be attempted. Here we illustrate only the hierarchical cluster analysis using UPGMA (Unweighted Pair-Group Method using Arithmetic averages) clustering.

In Table 11.5 of Mahanalobis' generalized distance between accessions, the smallest distance 6.909978 is between KKT and SSAT. We thus construct a sub-tree (of the dendrogram) by joining KKT and SSAT separated at a height 6.909978/2 = 3.45. Treating these two accessions as a single entity (say, C-1), the distance of other accessions with this will be worked out as follows:

Distance between C-1 and MVT = (distance between KKT and MVT + distance between SSAT and MVT) = (10.31317+ 31.6034)/2 = 20.96. Similarly other distances could be computed and new distance matrix is obtained (Table 11.6).

Table 11.6. Distance between cluster 1 (C-1) and other accessions

	C-1	MVT	NLAD	POLT
C-1	0			
MVT	20.95829	0		
NLAD	16.63491	54.26509	0	
POLT	60.4568	77.93759	36.2899	0

The smallest distance between accessions in Table 11.6 is 16.63 (between C-1 and NLAD). Hence, we amalgamate NLAD with C-1 at a height 16.63/2 = 8.31. The group C-1 and NLAD will be considered now as a single entity (C-2) and a new distance matrix is obtained. This process is continued until all the accessions are grouped into a single group, as shown below.

Among the distance between C-2 and other accessions (i.e. MVT and POLT), the smallest is between C-2 and MVT (37.61). Hence, MVT is amalgamated to C-2 at a height 37.61/2 = 18.85 and the newly formed group is called C-3.

The last accession (POLT) has a distance 63.15 with C-3 and joins at a height 63.15/2 = 31.57. Formation of clusters with respective distance and the resulting dendrogram is shown in Table 11.7 and Fig. 11.1.

Table 11.7. Formation of clusters and respective distances

Group	Elements of the group	Distance between elements
C-1	(KKT, SSAT)	3.45
C-2	(C-1, NLAD)	8.31
C-3	(C-2, MVT)	18.85
C-4	(C-3, PLOT)	31.57

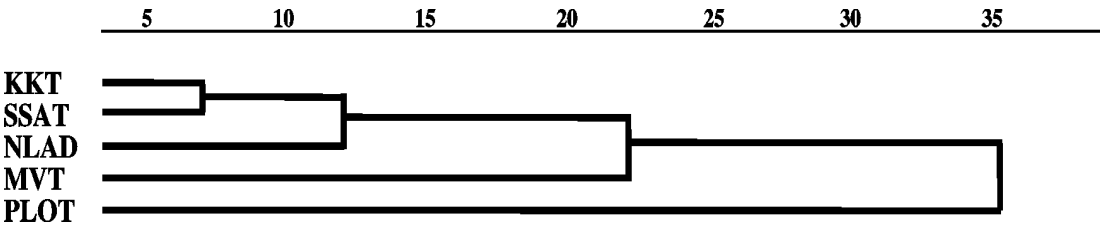


Figure 11.1. Average distance of clusters depicted in the form of dendrogram.

References

Cavalli-Sforza, L.L. and Edwards, A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21: 550-570.

Dillon, W.R. and Goldstein, M. 1984. Multivariate Analysis: Methods and Applications. Wiley, New York. 587p.

dos Sanots Dias, L.A., de Toledo Picoli, E.A., Rocha, R.B. and Alfenas, A.C. 2004. A priori choice of hybrid parents in plants. Genetics and Molecular Research 3(3): 356-368.

Johnson, R.A. and Wichern, D.W. 1992. Applied Multivariate Statistical Analysis, 3rd edition. Prentice-Hall, N.J. 642p.

Nei, M. 1972. Genetic distance between populations. Am. Natur. 106: 283-292.

Nei, M., Fuerst, P.A. and Chakraborty, R. 1978. Subunit molecular weight and genetic variability of proteins in natural populations. Proceedings of National Academy of Science, USA. 75(7): 3359-3362.

Nei, M., Tajima, F. and Tateno, Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. J. Molec. Evol. 19: 153-170.

Rao, C.R. 1952. Advanced Statistical Methods in Biometric Research. Wiley, New York. 448p.

Weir, B.S. and Cockerham, C.C. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Appendix I: Introduction to R and its use to perform statistical analysis for data presented in this manual

Introduction

The objective of the appendices is to explain how the R statistical software can be used to perform the analyses presented in the manual. R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R is freely available at http://www.r-project.org/.

Data in the Tables of this manual have been copied into text files readable by R. The contents of these text files are printed at the end of respective appendices. A gray background is used for R commands and a frame with a white background for the results returned by R. It is easy to test the R commands by copying the text of the gray areas into the R console.

Before beginning the computations, some parameters of R have to be set.

Set working directory

```
setwd ("d:/stat/R")
```

This is only an example to input the path of the directory where the text files containing the data have been copied. Note that "/" must be used instead of "\" in the path.

Sometimes a command like this appears in the text:

```
options (digits=9)
```

This is used to set the number of digits for output in order to obtain approximately the same number of digits as in the computations in the corresponding chapter.

References

R Development Core Team. 2008. R : A language and environment for statistical computing. R Foundation for statistical computing. Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.r-project.org.

Appendix II: Sampling methods

Illustration of simple random sampling

Suppose we have to choose a sample of five coconut palms from a population size of 80. The first step is to fill a vector (say *coconuts*) with numbers from 1 to 80.

```
(coconuts <- 1:80)
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
[76] 76 77 78 79 80
```

Sampling without replacement

```
sample (coconuts,5,replace=F)
```

```
[1] 80 32  7 18 23
```

Sampling with replacement

```
sample (coconuts,5,replace=T)
```

```
[1] 59 79 40 47 79
```

Note that in that case "79" occurs twice.

Appendix III: Frequency distribution of observations

Frequency distribution of qualitative data

Load and display data

```
(fruit.shape <- read.table("03-3-1-fruitshape.txt",header=T))
```

	Class	Frequency
1	Round	169
2	Egg-shaped	61
3	Pear-shaped	62
4	Elliptic	22

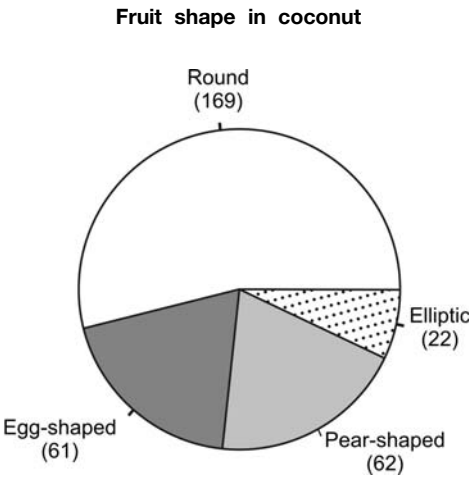
Percentages can be easily computed in order to obtain values of Table 3.1.

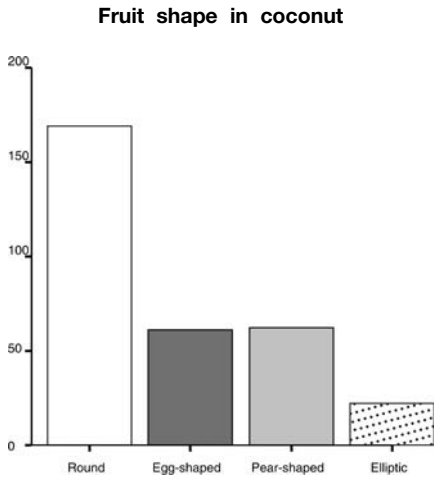
```
data.frame(fruit.shape,Percentage=round(fruit.shape$Frequency/sum (fruit.shape$
Frequency)*100,2))
```

	Class	Frequency	Percentage
1	Round	169	53.82
2	Egg-shaped	61	19.43
3	Pear-shaped	62	19.75
4	Elliptic	22	7.01

Plot pie chart and bar diagram

```
title <- "Fruit shape in coconut"
pie(fruit.shape$Frequency,labels=paste(fruit.shape$Class,'(',fruit.shape$Frequency,')'),
main=title)
barplot(fruit.shape$Frequency,names.arg=as.character(fruit.shape$Class),main=title,
ylim=c(0,200),axis.lty=1)
```





Frequency distribution of quantitative data

Load and display data

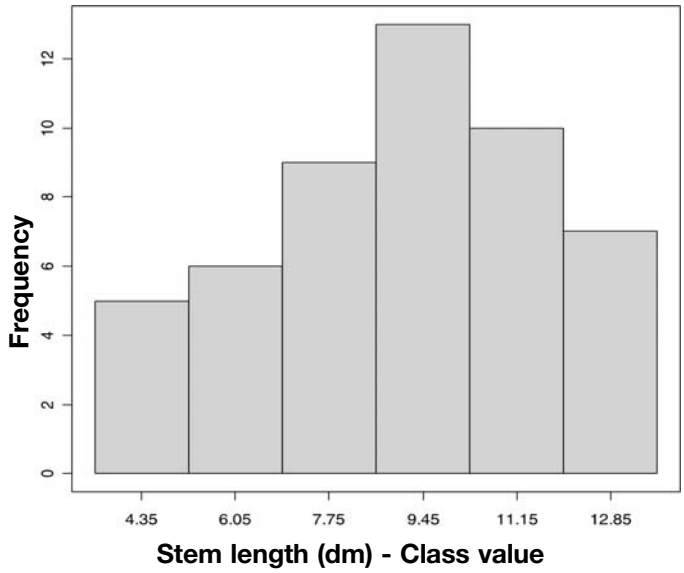
```
(stem.length <- scan("03-3-2-stemlength.txt"))
```

```
Read 50 items
 [1]  8.6  8.3  9.6  9.1 10.4  6.3 11.9  9.3  9.7  4.5  8.6  7.8  7.9 12.6  8.4
[16] 10.4 11.5 10.4  5.9  9.9  8.2  6.1  9.3 12.7 13.4 11.0  7.9  6.8  7.7  4.1
[31]  8.5 10.4  9.8  3.8  4.3 12.8  7.7  5.8 13.2  4.8 12.6  8.9 11.6 10.8  6.3
[46] 12.2  8.9 10.2  9.8 11.9
```

Frequency distribution and graphical representation

Although the class limits can be computed automatically by the *hist* function, in this example they are passed in the vector *breaks* in order to obtain the same histogram as in Figure 3.1.

```
breaks <- c(3.5,5.2,6.9,8.6,10.3,12,13.7)
stem.hist <- hist(stem.length,breaks=breaks,right=F,freq=T,col="lightgray",
                  xaxt="n",main=NULL,xlab="Stem length (dm) - Class value")
axis(side=1,at=stem.hist$mids,labels=stem.hist$mids)
box()
```



The values returned by *hist* in *stem.hist* can now be used in other functions. The following commands compute the data of Table 3.3.

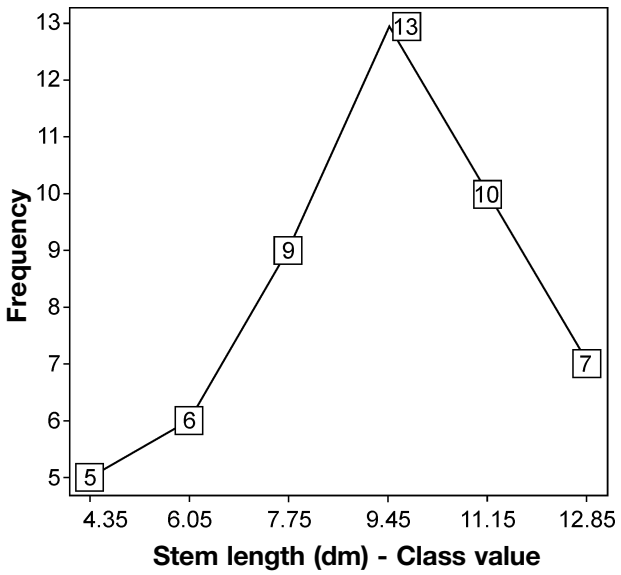
```
stem.cut <- cut(stem.length,breaks=breaks,right=F)
df    <- data.frame(
      Class.interval=levels(stem.cut),
      Class.value=stem.hist$mids,
      Frequency=stem.hist$counts,
      Cumul.freq=cumsum(stem.hist$counts),
      Percent=stem.hist$counts/sum(stem.hist$counts)*100,
      Cumul.percent=cumsum(stem.hist$counts)/sum(stem.hist$counts)*100)
df
```

	Class interval	Class value	Frequency	Cumul. Frequency	Percent	Cumul. percent
1	[3.5,5.2)	4.35	5	5	10	10
2	[5.2,6.9)	6.05	6	11	12	22
3	[6.9,8.6)	7.75	9	20	18	40
4	[8.6,10.3)	9.45	13	33	26	66
5	[10.3,12)	11.15	10	43	20	86
6	[12,13.7)	12.85	7	50	14	100

Upper limits of intervals are not included, so that these limits and the class values are slightly greater than in Table 3.3.

The following commands plot the frequency polygon of stem length. The plot is similar to Figure 3.2.

```
plot(stem.hist$mids,stem.hist$count,type="b",pch="",cex=3,lwd=2,
  xlab="Stem length (dm) - Class value",ylab="Frequency",axes=F)
text(stem.hist$mids,stem.hist$count,label=stem.hist$count)
axis(1,at=stem.hist$mids,labels=stem.hist$mids)
axis(2,4:14)
box()
```



Parameters of distributions

In the following we will consider only ungrouped data, since their management is not a problem in modern software and computations are more precise than with grouped data.

Measures of Central Tendency

```
c(Mean=mean(stem.length),Median=median(stem.length))
```

```
Mean  Median
9.052  9.200
```

Measures of Dispersion

```
range(stem.length)
```

```
[1] 3.8 13.4
```

As there is no function for mean deviation, we have first to define it.

```
mean.dev <- function(x){ sum(abs(x-mean(x)))/length(x) }
mean.dev(stem.length)
```

```
[1] 2.04592
```

We will use parameters h1, h2, h3, h4 and m1, m2, m3, m4 as described in this chapter. First we define the *h* function, then we call it to compute the parameters.

```
h <- function(x,n) { sum(x^n)/length(x) }
h1 <- h(stem.length,1)
h2 <- h(stem.length,2)
h3 <- h(stem.length,3)
h4 <- h(stem.length,4)
m1 <- h1
m2 <- h2-h1^2
m3 <- h3-3*h1*h2+2*h1^3
m4 <- h4-4*h1*h3+6*(h1^2)*h2-3*h1^4
c(m1,m2,m3,m4)
```

```
[1] 9.052000 6.373296 -4.446495 95.237575
```

We have now all the elements to compute the dispersion parameters.

```
result <- c(Mean=m1,Variance=m2,StDev=sqrt(m2),CV=sqrt(m2)/m1*100,
  Skewness=m3/m2^(3/2),Kurtosis=m4/m2^2-3)
round(result,3)
```

Mean	Variance	StDev	CV	Skewness	Kurtosis
9.052	6.373	2.525	27.889	-0.276	-0.655

Normal distribution

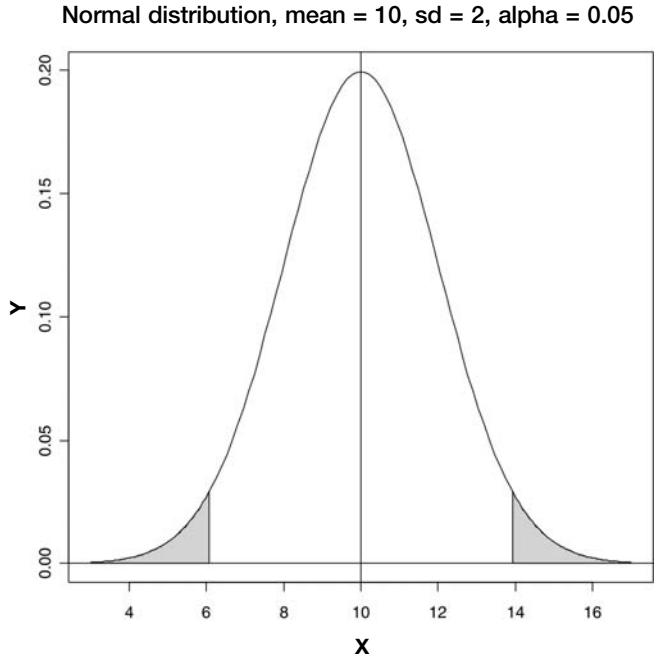
Normal curve

In order to plot the normal curve, we first load into R the *fill.area* function which will be called to fill some areas of the curve.

```
fill.area <- function(x1,x2,FUN,...){
  x <- seq(x1,x2,length=100)
  y <- FUN(x,...)
  xx <- c(x[1],x,x[length(x)])
  yy <- c(0,y,0)
  polygon(xx,yy,col="lightgray")
}
```

We can now plot the normal curve with parameters *mean* and *sd*. Another parameter, *alpha*, is such that the area shaded is *alpha*/2 for the left tail and for the right tail. Parameters *mean*, *sd*, and *alpha* can be changed.

```
mean <- 10; sd <- 2; alpha <- 0.05
x <- seq(mean-3.5*sd,mean+3.5*sd,length=100)
y <- dnorm(x,mean=mean,sd=sd)
plot(x,y,type="l",main=paste("Normal distribution",", mean =",mean,", sd =",sd,",
alpha =",alpha))
abline(h=0,v=mean)
q1 <- qnorm(alpha/2,mean=mean,sd=sd)
q2 <- qnorm(1-alpha/2,mean=mean,sd=sd)
fill.area(mean-3.5*sd,q1,dnorm,mean,sd)
fill.area(q2,mean+3.5*sd,dnorm,mean,sd)
```



As an example for a normal population of the stem length at 11 leaf scars of coconut palms with mean value 9.05 dm and standard deviation 2.5 dm, the probability of occurrence of stem lengths of 13.85 dm and above can be worked out as follows:

```
1-pnorm(13.85,mean=9.05,sd=2.5)
```

```
[1] 0.02742895
```

Confidence interval

Consider the data in *stem.length* as a sample drawn from a normal population with unknown mean and known standard deviation 2.5. The confidence interval of the population mean is computed as follows:

```
alpha <- 0.05
m <- mean(stem.length)
sigma.n <- 2.5/sqrt(length(stem.length))
round(c(m+qnorm(alpha/2)*sigma.n,m+qnorm(1-alpha/2)*sigma.n),3)
```

```
[1] 8.359 9.745
```

Minimum sample size for estimating the mean

The optimal sample size according to CV (%) and desired CI_{α} can be computed with the following *size* function.

```
size <- function(cv,b,alpha){ ceiling((qnorm(1-alpha/2)*cv/b)^2) }
size(20,7.5,0.05)
```

```
[1] 28
```

The following function calls *size* to comput a table of optimal sample size according to CV (%) and desired CI_{α}.

```
size.table <- function(cv,b,alpha){
  m <- sapply(cv,size,b,alpha)
  colnames(m) <- cv
  row.names(m) <- b
  m
}
```

It is now easy to print the values of Table 3.9.

```
size.table(cv=seq(5,25,2.5),b=seq(5,15,2.5),alpha=0.05)
```

	5	7.5	10	12.5	15	17.5	20	22.5	25
5	4	9	16	25	35	48	62	78	97
7.5	2	4	7	11	16	21	28	35	43
10	1	3	4	7	9	12	16	20	25
12.5	1	2	3	4	6	8	10	13	16
15	1	1	2	3	4	6	7	9	11

Contents of files used in the above computations and readable by R are printed below:

File 03-3-1-fruitshape.txt (data of Table 3.1)

Class	Frequency
Round	169
Egg-shaped	61
Pear-shaped	62
Elliptic	22

File 03-3-2-stemlength.txt (data of Table 3.2)

8.6	8.3	9.6	9.1	10.4	6.3	11.9	9.3	9.7	4.5	8.6	7.8	7.9	12.6	8.4	10.4	11.5
10.4	5.9	9.9	8.2	6.1	9.3	12.7	13.4	11.7	7.9	6.8	7.7	4.1	8.5	10.4	9.8	3.8
4.3	12.8	7.7	5.8	13.2	4.8	12.6	8.9	11.6	10.8	6.3	12.2	8.9	10.2	9.8	11.9	

Appendix IV: Estimation and tests of significance

Estimation, t-test for means

Load and display data

```
(wct <- scan("04-wct.txt"))
```

```
Read 26 items
 [1] 177.25 154.50 173.25 193.50 227.50 155.25 168.00 233.00 150.00 158.75
[11] 230.00 200.75 169.75 176.75 158.00 164.25 154.50 162.50 186.50 207.00
[21] 250.00 157.50 228.50 216.50 227.50 181.50
```

Compute mean, variance, t, one-tailed and two-tailed test

```
n <- length(wct)
c(mean=mean(wct),s2=sd(wct)^2,t=(mean(wct)-
172)/sd(wct)*sqrt(n),t1tail=qt(0.95,n-1),t2tail=qt(0.975,n-1))
```

mean	s2	t	t1tail	t2tail
187.0192	945.3296	2.4908	1.7081	2.0595

One-tailed and two-tailed test can also be computed with the *t.test* function.

```
t.test(wct,mu=172)
```

```
    One Sample t-test

data: wct
t = 2.4908, df = 25, p-value = 0.01975
alternative hypothesis: true mean is not equal to 172
95 percent confidence interval:
 174.60 199.44
sample estimates:
mean of x
 187.02
```

```
t.test(wct,mu=172,alternative="greater")
```

```
   One Sample t-test

data: wct
t = 2.4908, df = 25, p-value = 0.009873
alternative hypothesis: true mean is greater than 172
95 percent confidence interval:
 176.72 Inf
sample estimates:
mean of x
 187.02
```

We reject the null hypothesis at level 0.02 for two-tailed test and 0.0099 for one-tailed test.

Comparison of two sample means (independent samples)

Load and display data

```
(twopop <- read.table("04-1-twopop.txt",header=T))
```

	pop	nuts
1	monocrop	16.30
2	monocrop	15.50
3	monocrop	27.30
4	monocrop	22.60
5	monocrop	12.20
6	monocrop	18.70
7	monocrop	7.25
8	monocrop	9.70
9	monocrop	21.30
10	monocrop	15.50
11	monocrop	22.20
12	monocrop	13.20
13	monocrop	19.00
14	monocrop	17.40
15	monocrop	28.80
16	monocrop	14.90
17	intercrop	21.40
18	intercrop	13.20
19	intercrop	26.80
20	intercrop	29.30
21	intercrop	17.40
22	intercrop	16.30
23	intercrop	12.10
24	intercrop	9.00

```
25    intercrop    20.80
26    intercrop    17.70
27    intercrop    19.40
28    intercrop    15.20
29    intercrop    18.30
30    intercrop    18.00
31    intercrop    25.40
32    intercrop    17.30
33    intercrop    18.80
34    intercrop    19.50
```

Two-tailed test

```
monocrop <- twopop$nuts[twopop$pop==”monocrop”]
intercrop <- twopop$nuts[twopop$pop==”intercrop”]
t.test(monocrop,intercrop,var.equal=T)
```

```
    Two Sample t-test

data: monocrop and intercrop
t = -0.5598, df = 32, p-value = 0.5795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.8494 2.7585
sample estimates:
mean of x mean of y
 17.616 18.661
```

One-tailed test

```
t.test(monocrop,intercrop,var.equal=T,alternative=”less”)
```

```
    Two Sample t-test

data: monocrop and intercrop
t = -0.5598, df = 32, p-value = 0.2897
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 2.1178
sample estimates:
mean of x mean of y
 17.616 18.661
```

In both cases, we do not reject the null hypothesis.

Comparison of two related sample means (matched pairs)

Load and display data

```
(pairs <- read.table("04-2-pairs.txt",header=T))
```

	Pre.treatment	Post.treatment
1	16.30	21.4
2	15.50	13.2
3	27.30	26.8
4	22.60	29.3
5	12.20	17.4
6	18.70	16.3
7	7.25	12.1
8	9.70	9.0
9	21.30	20.8
10	15.50	17.7
11	22.20	19.4
12	13.20	15.2
13	19.00	18.3
14	17.40	18.0
15	28.80	25.4
16	14.90	17.3

Compute t-test

```
t.test(pairs$Pre.treatment,pairs$Post.treatment,paired=T)
```

```
    Paired t-test

data: pairs$Pre.treatment and pairs$Post.treatment
t = -1.2335, df = 15, p-value = 0.2364
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.68536  0.71661
sample estimates:
mean of the differences
 -0.98437
```

F- test for equality of two variances

```
var.test(monocrop,intercrop)
```

```
   F test to compare two variances

data: monocrop and intercrop
F = 1.3764, num df = 15, denom df = 17, p-value = 0.5228
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.50547 3.87155
sample estimates:
ratio of variances
 1.3764
```

Chi-square test of significance for goodness of fit for frequency data

Load data

```
x <- c(46,47,48,48,48,49,52,52,54,54,55,57,65,65,66,67,69,72,73,76)
```

Load the *fit.test* function

```
fit.test <- function(x,nclass,alpha=0.05){
 options(digits=4)
 breaks <- c(min(x),qnorm((1:(nclass-1))/nclass,mean=60,sd=10),max(x))
 observed <- hist(x,breaks=breaks,plot=F)$count
 expected <- length(x)/(length(breaks)-1)
print(t(data.frame(observed,expected,row.names=levels(cut(x,breaks,include.
lowest= T)))))
c(chi2=sum(((observed-expected)^2)/expected),chi2.0=qchisq(1-
alpha,length(breaks)-2),df=length(breaks)-2)
}
```

Compute Chi-Square Test (4 classes)

```
fit.test(x,nclass=4)
```

	[46,53.3]	(53.3,60]	(60,66.7]	(66.7,76]
observed	8	4	3	5
expected	5	5	5	5

chi2	chi2.0	df
2.800	7.815	3.000

The null hypothesis is not rejected. However the test is very sensitive to the number of classes, as shown in the next example.

Compute Chi-Square Test (3 classes)

```
fit.test(x,nclass=3)
```

	[46,55.7]	(55.7,64.3]	(64.3,76]
observed	11.000	1.000	8.000
expected	6.667	6.667	6.667

chi2	chi2.0	df
7.900	5.991	2.000

With 3 classes, the null hypothesis is rejected.

Chi-square test for independence

Load and display data

```
(embryos <- matrix(c(28,39,45,32,18,17),byrow=T,nrow=2))
```

	[,1]	[,2]	[,3]
[1,]	28	39	45
[2,]	32	18	17

Compute Chi-Square Test

```
chisq.test(embryos)
```

```
        Pearson's Chi-squared test

data: embryos
X-squared = 9.966, df = 2, p-value = 0.006855
```

The null hypothesis is rejected.

Chi-square test for homogeneity of variances (Bartlett test)

Load and display data

```
(fourpop <- read.table("04-5-fourpop.txt",header=T))
```

	accession	weight
1	1	438
2	1	449
3	1	453
4	1	518
5	1	564
6	1	608
7	1	610
8	1	651
9	1	680
10	1	700
11	2	1004
12	2	1018
•		
•		
•		
48	4	824
49	4	838

(See complete data below)

Compute Chi-Square Test

```
bartlett.test(weight~accession,data=fourpop)
```

```
    Bartlett test for homogeneity of variances

data: weight by accession
Bartlett's K-squared = 37.24, df = 3, p-value = 4.084e-08
```

The hypothesis of variance equality is rejected.

Contents of files used in the above computations and readable by R are printed below:

File 04-wct.txt

```
177.25 154.50 173.25 193.50 227.50 155.25 168.00 233.00 150.00 158.75 230.00
200.75 169.75 176.75 158.00 164.25 154.50 162.50 186.50 207.00 250.00 157.50
228.50 216.50 227.50 181.50
```

File 04-1-twopop.txt (data of Table 4.1)

pop	nuts
monocrop	16.3
monocrop	15.5
monocrop	27.3
monocrop	22.6
monocrop	12.2
monocrop	18.7
monocrop	7.25
monocrop	9.70
monocrop	21.3
monocrop	15.5
monocrop	22.2
monocrop	13.2
monocrop	19.0
monocrop	17.4
monocrop	28.8
monocrop	14.9
intercrop	21.4
intercrop	13.2
intercrop	26.8
intercrop	29.3
intercrop	17.4
intercrop	16.3
intercrop	12.1
intercrop	9.0
intercrop	20.8
intercrop	17.7
intercrop	19.4
intercrop	15.2
intercrop	18.3
intercrop	18.0
intercrop	25.4
intercrop	17.3
intercrop	18.8
intercrop	19.5

File 04-2-pairs.txt (data of Table 4.2)

Pre.treatment	Post.treatment
16.3	21.4
15.5	13.2
27.3	26.8
22.6	29.3
12.2	17.4
18.7	16.3
7.25	12.1
9.7	9.0
21.3	20.8
15.5	17.7
22.2	19.4
13.2	15.2
19.0	18.3
17.4	18.0
28.8	25.4
14.9	17.3

File 04-5-fourpop.txt (data of Table 4.3)

accession	weight
1	438
1	449
1	453
1	518
1	564
1	608
1	610
1	651
1	680
1	700
2	1004
2	1018
2	1019
2	1032
2	1045
2	1053
2	1056
2	1060
2	1068
2	1074

2	1087
2	1095
2	1116
2	1141
3	1270
3	1421
3	1425
3	1435
3	1445
3	1446
3	1461
3	1506
3	1526
3	1568
3	1610
3	1780
4	770
4	775
4	784
4	786
4	788
4	790
4	791
4	795
4	802
4	806
4	813
4	824
4	838

Appendix V: Analysis of relationship between variables

Correlation

Load and display data

```
(fruits <- read.table("05-2-fruits.txt",header=T))
```

	FW	NW	VC	KW	CW
1	1216	662	180	346	172
2	1445	735	200	383	187
3	786	466	110	262	157
4	784	467	110	272	152
5	750	464	120	262	155
6	1004	638	190	305	194
7	838	505	140	279	170
8	892	560	180	264	165
9	1019	614	190	321	198
10	860	486	170	252	158
11	1060	701	230	362	224
12	928	569	180	305	194
13	1568	875	310	429	245
14	1461	868	300	414	250
15	1141	686	270	386	209
16	1170	722	230	400	206
17	960	548	140	275	162
18	712	437	120	240	144
19	1002	532	130	280	174
20	1183	555	110	286	164

Compute the correlation matrix

```
round(m <- cor(fruits),3)
```

	FW	NW	VC	KW	CW
FW	1.000	0.929	0.769	0.888	0.772
NW	0.929	1.000	0.922	0.960	0.924
VC	0.769	0.922	1.000	0.896	0.929
KW	0.888	0.960	0.896	1.000	0.894
CW	0.772	0.924	0.929	0.894	1.000

This matrix is identical to Table 5.4.

Testing the equality of two correlation coefficients

Since the statistic is $(z_i - z_j)/\sqrt{var(z_i - z_j)}$, a simple procedure is to compute the matrix of $z_i/\sqrt{(2var\ z_i)}$.

```
z <- function(r) { 0.5*log((1+r)/(1-r)) }
(z.m <- z(m)/sqrt(2/(nrow(fruits)-3)))
```

	FW	NW	VC	KW	CW
FW	Inf	4.823545	2.964272	4.114632	2.988287
NW	4.823545	Inf	4.673527	5.662620	4.705543
VC	2.964272	4.673527	Inf	4.231079	4.815294
KW	4.114632	5.662620	4.231079	Inf	4.206436
CW	2.988287	4.705543	4.815294	4.206436	Inf

The statistics are then the differences between appropriate values of this matrix.

```
z.m["NW","CW"]-z.m["FW","CW"]
```

```
[1] 1.717256
```

Since this difference is lower than 1.96, we do not reject the null hypothesis at 5% level.

Simple linear regression

Compute the regression

```
(reg <- lm(NW~FW,data=fruits))
```

```
Call:
lm(formula = NW ~ FW, data = fruits)

Coefficients:
(Intercept)          FW
    94.1086      0.4913
```

Analysis of variance

```
anova(reg)
```

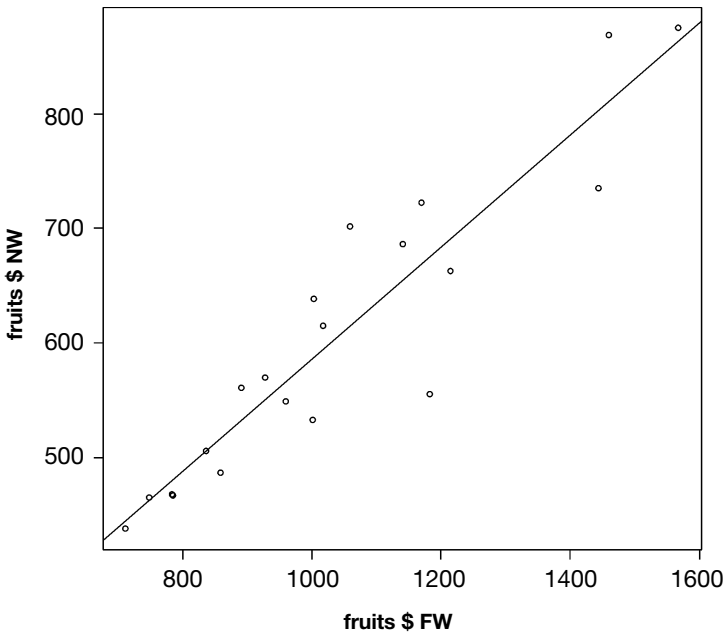
```
Analysis of Variance Table

Response: NW
           Df  Sum Sq Mean Sq F value    Pr(>F)
FW          1  274619  274619  114.27 3.170e-09 ***
Residuals  18   43260    2403
—
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We reject the null hypopthesis since the F value (114.27) is highly significant.

Plot the points and regression line

```
plot(fruits$FW,fruits$NW)
abline(reg)
```



Multiple linear regression

Compute the regression and display the results

```
reg <- lm(CW~FW+NW+VC+KW,data=fruits)
summary(reg)
```

```
Call:
lm(formula = CW ~ FW + NW + VC + KW, data = fruits)

Residuals:
Min                1Q          Median          3Q          Max
-17.139         -6.489          -1.573          6.736       13.498

Coefficients:
                Estimate    Std. Error      t value     Pr(>|t|)
(Intercept)     59.06062    17.55228        3.365       0.00425  **
FW              -0.06129    0.03281         -1.868      0.08136
NW              0.26830     0.11631         2.307       0.03575  *
VC              0.11014     0.12429         0.886       0.38950
KW              0.02072     0.13922         0.149       0.88365
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.12 on 15 degrees of freedom
Multiple R-Squared: 0.9134,         Adjusted R-squared: 0.8903
F-statistic: 39.56 on 4 and 15 DF,  p-value: 8.434e-08
```

The cofficients and tests are identical to those of table 5.5d.

Path-Coefficient Analysis

The path coefficients can be obtained by calling the *path.coeff* function, which has to be loaded into R.

```
path.coeff <- function(model,data){
  reg <- lm(model,data)
  dim <- length(reg$coeff)-1
  m.cor <- cor(reg$model[-1])
  f <- function(i) reg$coeff[i+1]*sd(reg$model[,i+1]/sd(reg$model[,1]))
  b <- sapply(1:dim,f)
  t(m.cor*b)
}
```

Compute the path coefficients

```
path.coeff(CW~FW+NW+VC+KW,data=fruits)
```

	FW	NW	VC	KW
FW	-0.4908291	1.055459	0.1713247	0.03592978
NW	-0.4562096	1.135553	0.2055627	0.03884139
VC	-0.3772189	1.047116	0.2229240	0.03626035
KW	-0.4357440	1.089805	0.1997264	0.04047188

This table is identical to Table 5.7.

Contents of files used in the above computations and readable by R are printed below:

File 05-2-fruits.txt (data of Table 5.2)

FW	NW	VC	KW	CW
1216	662	180	346	172
1445	735	200	383	187
786	466	110	262	157
784	467	110	272	152
750	464	120	262	155
1004	638	190	305	194
838	505	140	279	170
892	560	180	264	165
1019	614	190	321	198
860	486	170	252	158
1060	701	230	362	224
928	569	180	305	194
1568	875	310	429	245
1461	868	300	414	250
1141	686	270	386	209
1170	722	230	400	206
960	548	140	275	162
712	437	120	240	144
1002	532	130	280	174
1183	555	110	286	164

Appendix VI: Basic principles for planning and conducting coconut field trials

Randomization

Consider an experiment having four treatments 1, 2, 3, and 4 that are to be randomly allocated among 24 experimental units (say, trees) so that each treatment is replicated six times.

Complete randomization of the treatments

```
sample(rep(1:4,each=6))
```

```
[1] 1 2 1 4 1 4 1 4 3 1 3 2 3 4 3 4 2 2 4 2 1 2 3 3
```

The obtained sequence of treatments can be applied in this order to the experimental units. Now we consider the same treatments applied to 24 experimental units grouped in 6 blocks of 4 units.

Randomization of the treatments within blocks

```
sapply(1:6,function(i) sample(1:4))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	2	1	1	2	1	4
[2,]	3	3	3	1	3	3
[3,]	4	4	4	3	2	1
[4,]	1	2	2	4	4	2

Each column of the matrix corresponds to a block and contains the 4 treatments.

Appendix VII: Basic experimental designs for coconut trials

In the following, we use the function aov to perform the analyses of variance because data are balanced for all the experiments. In case of unbalanced data, the function lm would be appropriate.

Completely Randomized Design (CRD)

Load and display data

```
(crd <- read.table("07-2-complete-random.txt",header=T))
```

	treatment	percent
1	T1	30.3
2	T1	28.6
3	T1	26.6
4	T1	33.4
5	T1	34.4
6	T1	29.7
7	T2	37.0
8	T2	34.7
9	T2	41.5
10	T2	36.5
11	T2	38.1
12	T2	35.9
13	T3	-10.2
14	T3	-5.3
15	T3	-13.3
16	T3	-6.8
17	T3	-18.1
18	T3	-22.1
19	T4	-45.3
20	T4	-19.8
21	T4	-9.6
22	T4	-28.9
23	T4	-49.6
24	T4	-35.1

Perform the analysis and display the analysis of variance table

```
options(digits=9)
crd.aov <- aov(percent~treatment,data=crd)
anova(crd.aov)
```

```
Analysis of Variance Table

Response: percent
             Df     Sum Sq    Mean Sq   F value     Pr(>F)
Treatment     3 19941.588  6647.196  92.40495 6.825e-12 ***
Residuals    20  1438.710    71.936

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 7.4. The p-value Pr(>F) is given directly, showing that the treatment effects are highly significantly different, so that the tabular F value is not needed. However, it is easy to obtain this value ($F_{0.05}$ with 3 and 20 df):

Compute the tabular F value

```
qf(0.95,3,20)
```

```
[1] 3.09839121
```

This value is the same as in Table 7.4.

Compute the means

```
model.tables(crd.aov,type="means")
```

```
Tables of means
Grand mean

5.94166667

 treatment
treatment
    T1     T2     T3     T4
 30.50  37.28 -12.63 -31.38
```

The means of the treatments are the same as in Table 7.2.

R provides several methods of multiple comparisons. One of them is based on the Studentized range statistic, Tukey's 'Honest Significant Difference' method.

Compute Tukey's Honest Significant Difference

```
TukeyHSD(crd.aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percent ~ treatment, data = crd)

$treatment
             diff        lwr        upr
T2-T1   6.78333333 -6.92245426 20.48912093
T3-T1 -43.13333333 -56.83912093 -29.42754574
T4-T1 -61.88333333 -75.58912093 -48.17754574
T3-T2 -49.91666667 -63.62245426 -36.21087907
T4-T2 -68.66666667 -82.37245426 -54.96087907
T4-T3 -18.75000000 -32.45578759  -5.04421241
```

The results are the differences (diff) between means of treatments along with the lower (lwr) and upper (upr) bounds. The difference is significant if the interval [lwr,upr] does not overlap zero. Here all the differences are significant excepted T2-T1, so that we can order the treatment means as follows:

T4 < T3 < T1 = T2

Randomized Complete Block Design (RCBD)

Load and display data

```
(rcbd <- read.table("07-6-random-block.txt",header=T))
```

	cultivar	block	nuts
1	AOT	B1	74.95
2	AGT	B1	80.18
3	PHOT	B1	70.91
4	FMS	B1	65.49
5	SSG	B1	93.80
6	FJT	B1	69.26
7	CCT	B1	90.83

8	JGT	B1	71.11
9	LCT	B1	120.51
10	AOT	B2	54.51
11	AGT	B2	71.13
12	PHOT	B2	61.45
13	FMS	B2	55.63
14	SSG	B2	77.65
15	FJT	B2	51.01
16	CCT	B2	78.75
17	JGT	B2	80.13
18	LCT	B2	79.80
19	AOT	B3	62.60
20	AGT	B3	77.80
21	PHOT	B3	68.80
22	FMS	B3	58.70
23	SSG	B3	82.60
24	FJT	B3	61.70
25	CCT	B3	85.80
26	JGT	B3	74.60
27	LCT	B3	98.70

Perform the analysis and display the analysis of variance table

```
rcbd.aov <- aov(nuts~cultivar+block,data=rcbd)
(rcbd.anova <- anova(rcbd.aov))
```

```
Analysis of Variance Table
Response: nuts
           Df  Sum Sq Mean Sq F value    Pr(>F)
Cultivar    8 4267.762 533.470 12.01855 1.8870e-05 ***
Block       2  896.148 448.074 10.09466 0.0014599 **
Residuals  16  710.196  44.387

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 7.8. The p-value Pr(>F) are given directly, showing that the treatment and block effects are significantly different.

Compute the means

```
(rcbd.tables <- model.tables(rcbd.aov,type="means"))
```

```
Tables of means
Grand mean

74.7555556

cultivar
cultivar
   AGT    AOT    CCT    FJT    FMS    JGT    LCT   PHOT    SSG
 76.37  64.02  85.13  60.66  59.94  75.28  99.67  67.05  84.68

  block
  block
    B1     B2     B3
 81.89  67.78  74.59
```

The means of the cultivars and of the blocks are the same as in Table 7.6.

Mean comparisons

We first compute the values necessary for comparing the means:

```
sigma2 <- rcbd.anova[[3]][3]
r <- rcbd.anova$Df[2]+1 # Number of replicates
t0 <- qt(0.975,rcbd.aov$df.residual)
cd <- t0*sqrt(2*sigma2/r)
print(c(sigma2,r,t0,cd),digits=5)
```

```
[1] 44.3873  3.0000  2.1199 11.5319
```

As we have now the CD (11.5319) we could obtain the groups manually. However we can load the following function into R in order to print them automatically:

```
groups <- function(x,cd){
  x <- sort(x)
  n <- length(x)
  m <- sapply(1:n,function(i) { j <- i:n ; max(j[x[j]<=x[i]+cd]) })
  group.min <- sapply(split(1:n,m),min)
  group.max <- as.numeric(names(group.min))
  ngroup <- length(group.min)
  group <- matrix(" ",nrow=n,ncol=ngroup)
  for (j in 1:ngroup) group[group.min[j]:group.max[j],j] <- letters[j]
  data.frame(x,groups=apply(group,1,paste,collapse=""))
}
```

We can now easily print the groups:

```
print(groups(rcbd.tables$tables$cultivar,cd),digits=4)
```

	x	groups
FMS	59.94	a
FJT	60.66	a
AOT	64.02	ab
PHOT	67.05	abc
JGT	75.28	bcd
AGT	76.37	cd
SSG	84.68	d
CCT	85.13	d
LCT	99.67	e

Latin Square Design (LSD)

Load and display data

```
(lsd <- read.table("07-9-latin-square.txt",header=T))
```

	person	day	cultivar	percent
1	Technician-1-AM	Day 1	WCT	25
2	Technician-2-AM	Day 1	WAT	10
3	Technician-1-PM	Day 1	PHOT	85
4	Technician-2-PM	Day 1	LCT	65
5	Technician-1-AM	Day 2	WAT	25
6	Technician-2-AM	Day 2	WCT	40
7	Technician-1-PM	Day 2	LCT	70
8	Technician-2-PM	Day 2	PHOT	75
9	Technician-1-AM	Day 3	PHOT	80
10	Technician-2-AM	Day 3	LCT	65
11	Technician-1-PM	Day 3	WAT	20
12	Technician-2-PM	Day 3	WCT	45
13	Technician-1-AM	Day 4	LCT	55
14	Technician-2-AM	Day 4	PHOT	85
15	Technician-1-PM	Day 4	WCT	30
16	Technician-2-PM	Day 4	WAT	20

Perform the analysis and display the analysis of variance table

```
lsd.aov <- aov(percent~person+day+cultivar,data=lsd)
anova(lsd.aov)
```

```
Analysis of Variance Table

Response: percent
            Df    Sum Sq   Mean Sq  F value     Pr (>F)
Person       3    67.187    22.396  0.37391  0.77529716
Day          3   129.688    43.229  0.72174  0.57468252
Cultivar     3  9467.187  3155.729 52.68696 0.00010552 ***
Residuals    6   359.375    59.896

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 7.11. The p-value Pr(>F) are given directly, showing that only the cultivar effects are significantly different.

Compute the means

```
model.tables(lsd.aov,type="means")
```

```
Tables of means
Grand mean

49.6875

 person
person
Technician-1-AM Technician-1-PM Technician-2-AM Technician-2-PM
          46.25           51.25           50.00           51.25

 day
day
 Day 1   Day2   Day 3   Day 4
 46.25  52.50  52.50  47.50

 cultivar
cultivar
   LCT   PHOT    WAT    WCT
 63.75  81.25  18.75  35.00
```

The means are computed for person, day, and cultivar (in Table 7.10 either the means or the sums are computed, according to the factor).

Compute Tukey's Honest Significant Difference for cultivars

```
TukeyHSD(lsd.aov,"cultivar")
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percent ~ person + day + cultivar, data = lsd)

$cultivar
                diff        lwr         upr
PHOT-LCT       17.50 -1.44410813  36.44410813
WAT-LCT       -45.00 -63.94410813 -26.05589187
WCT-LCT       -28.75 -47.69410813  -9.80589187
WAT-PHOT      -62.50 -81.44410813 -43.55589187
WCT-PHOT      -46.25 -65.19410813 -27.30589187
WCT-WAT        16.25 -2.69410813  35.19410813
```

According to theses results, the cultivars can be ordered as follows:

$$WAT = WCT < LCT = PHOT$$

Contents of files used in the previous computations and readable by R are printed below:

File 07-2-complete-random.txt (data of Table 7.2)

```
treatment percent
T1           30.3
T1           28.6
T1           26.6
T1           33.4
T1           34.4
T1           29.7
T2           37.0
T2           34.7
T2           41.5
T2           36.5
T2           38.1
T2           35.9
T3          -10.2
T3           -5.3
T3          -13.3
T3           -6.8
```

T3	-18.1
T3	-22.1
T4	-45.3
T4	-19.8
T4	-9.6
T4	-28.9
T4	-49.6
T4	-35.1

File 07-6-random-block.txt (data of Table 7.6)

cultivar	block	nuts
AOT	B1	74.95
AGT	B1	80.18
PHOT	B1	70.91
FMS	B1	65.49
SSG	B1	93.80
FJT	B1	69.26
CCT	B1	90.83
JGT	B1	71.11
LCT	B1	120.51
AOT	B2	54.51
AGT	B2	71.13
PHOT	B2	61.45
FMS	B2	55.63
SSG	B2	77.65
FJT	B2	51.01
CCT	B2	78.75
JGT	B2	80.13
LCT	B2	79.80
AOT	B3	62.60
AGT	B3	77.80
PHOT	B3	68.80
FMS	B3	58.70
SSG	B3	82.60
FJT	B3	61.70
CCT	B3	85.80
JGT	B3	74.60
LCT	B3	98.70

File 07-9-latin-square.txt (data of Table 7.9)

person	day	cultivar	percent
Technician-1-AM	"Day 1"	WCT	25
Technician-2-AM	"Day 1"	WAT	10
Technician-1-PM	"Day 1"	PHOT	85
Technician-2-PM	"Day 1"	LCT	65
Technician-1-AM	"Day 2"	WAT	25
Technician-2-AM	"Day 2"	WCT	40
Technician-1-PM	"Day 2"	LCT	70
Technician-2-PM	"Day 2"	PHOT	75
Technician-1-AM	"Day 3"	PHOT	80
Technician-2-AM	"Day 3"	LCT	65
Technician-1-PM	"Day 3"	WAT	20
Technician-2-PM	"Day 3"	WCT	45
Technician-1-AM	"Day 4"	LCT	55
Technician-2-AM	"Day 4"	PHOT	85
Technician-1-PM	"Day 4"	WCT	30
Technician-2-PM	"Day 4"	WAT	20

Appendix VIII: Experimental designs for coconut trials with modified blocking

Balanced Incomplete Block Design (BIBD)

Load and display data

```
bibd <- read.table("08-1-bibd.txt",header=T)
bibd$block <- as.factor(bibd$block)
bibd
```

	block	treatment	seedlings
1	1	P7	40
2	1	P5	55
3	1	P4	65
4	2	P3	72
5	2	P5	58
6	2	P6	25
7	3	P5	63
8	3	P2	58
9	3	P1	67
10	4	P9	41
11	4	P3	80
12	4	P4	61
13	5	P5	52
14	5	P8	71
15	5	P9	49
16	6	P8	78
17	6	P7	46
18	6	P6	33
19	7	P8	69
20	7	P3	71
21	7	P2	61
22	8	P9	38
23	8	P1	70
24	8	P6	36
25	9	P9	34
26	9	P7	41
27	9	P2	52
28	10	P2	58

29	10	P4	68
30	10	P6	41
31	11	P3	74
32	11	P7	44
33	11	P1	71
34	12	P8	77
35	12	P1	61
36	12	P4	68

Perform the analysis and display the analysis of variance table

The first line defines the contrasts necessary for the computation of the adjusted means. Note that the order of the terms in the model *seedlings~block+treatment* is important: the block sum of squares is computed first, then the adjusted treatment sum of squares is computed. The model *seedlings~ treatment+ block* would give a different analysis.

```
options(contrasts=c("contr.sum","contr.poly"))
bibd.aov <- aov(seedlings~block+treatment,data=bibd)
(bibd.anova <- anova(bibd.aov))
```

```
Analysis of Variance Table

Response: seedlings
           Df  Sum Sq Mean Sq F value    Pr(>F)
Block      11 2037.556 185.232 7.77501 0.00015680 ***
Treatment   8 5254.815 656.852 27.57093 5.9991e-08 ***
Residuals  16  381.185  23.824

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 8.3. The p-value Pr(>F) is given directly, showing that the treatment and block effects are highly significantly different.

Compute the adjusted means

The effects of the factors are stored in *bibd.aov* and can be easily recovered:

```
(co <- coef(bibd.aov))
```

(Intercept)	block1	block2	block3	block4
56.888888889	-2.074074074	-3.555555556	2.333333333	0.222222222
block5	block6	block7	block8	block9
-0.259259259	2.444444444	-1.629629630	1.370370370	-5.111111111
block10	block11	treatment1	treatment2	treatment3
4.074074074	1.518518519	8.888888889	0.444444444	18.222222222
treatment4	treatment5	treatment6	treatment7	treatment8
7.888888889	1.000000000	-24.222222222	-13.333333333	16.555555556

Note that only 11 block effects and 8 treatment effects are computed. Because we have defined the contrasts as "*contr.sum*", the 9 treatment effects sum to 0 so that the 9th treatment effect is minus the sum of the first 8. The adjusted means can then be computed as follows, adding the intercept *co[1]* to the treatment effects:

```
co.treatment <- co[bibd.aov$assign==2]
effects.treatment <- c(co.treatment,-sum(co.treatment))
adjusted.means <- co[1]+effects.treatment
names(adjusted.means) <- levels(bibd$treatment)
print(sort(adjusted.means),digits=5)
```

P6	P9	P7	P2	P5	P4	P1	P8	P3
32.667	41.444	43.556	57.333	57.889	64.778	65.778	73.444	75.111

The means of the treatments are the same as in Table 8.4.

Mean comparisons

We first compute the values necessary for comparing the means:

```
sigma2 <- bibd.anova[[3]][3]
b <- bibd.anova$Df[1]+1 # Number of blocks
v <- bibd.anova$Df[2]+1 # Number of treatments
r <- nrow(bibd)/v # Number of replicates
k <- nrow(bibd)/b
lambda <- r*(k-1)/(v-1) # Not necessarily 1
t0 <- qt(0.975,bibd.aov$df.residual)
cd <- t0*sqrt(2*k*sigma2/lambda/v)
print(c(sigma2,b,v,r,k,lambda,t0,cd),digits=5)
```

```
[1] 23.8241 12.0000  9.0000  4.0000  3.0000  1.0000  2.1199  8.4485
```

As we have now the CD (8.4485) we could obtain the groups manually. However we can load the following function into R in order to print them automatically:

```
groups <- function(x,cd){
   x <- sort(x)
   n <- length(x)
   m <- sapply(1:n,function(i) { j <- i:n ; max(j[x[j]<=x[i]+cd]) })
   group.min <- sapply(split(1:n,m),min)
   group.max <- as.numeric(names(group.min))
   ngroup <- length(group.min)
   group <- matrix(" ",nrow=n,ncol=ngroup)
   for (j in 1:ngroup) group[group.min[j]:group.max[j],j] <- letters[j]
   data.frame(x,groups=apply(group,1,paste,collapse=""))
}
```

We can now easily print the groups:

```
print(groups(adjusted.means,cd),digits=5)
```

	x	groups
P6	32.667	a
P9	41.444	b
P7	43.556	b
P2	57.333	c
P5	57.889	c
P4	64.778	c
P1	65.778	cd
P8	73.444	de
P3	75.111	e

Augmented Block Designs

Load and display data

```
augmented <- read.table("08-1-augmented.txt",header=T)
augmented$block <- as.factor(augmented$block)
augmented
```

	block	treatment	type	wax
1	1	H8	test	74
2	1	LCT x GBGD	check	78
3	1	WCT	check	78
4	1	H3	test	70

5	1	WCT x COD	check	83
6	1	COD x WCT	check	77
7	1	H7	test	75
8	2	WCT	check	91
9	2	COD x WCT	check	81
10	2	WCT x COD	check	79
11	2	LCT x GBGD	check	81
12	2	H1	test	79
13	2	H5	test	78
14	3	H4	test	96
15	3	LCT x GBGD	check	87
16	3	WCT x COD	check	92
17	3	H2	test	89
18	3	WCT	check	81
19	3	COD x WCT	check	79
20	3	H6	test	82

Note that a column *type* has been included in the data in order to distinguish between test and check treatments.

In order to perform the analysis, we need that the check levels of the factor *treatment* appear first in the list of levels. Unfortunately, the levels appear naturally in alphabetical order:

```
levels (augmented$treatment)
```

```
 [1] "COD x WCT"   "H1"        "H2"   "H3"   "H4"
 [6] "H5"          "H6"        "H7"   "H8"   "LCT x GBGD"
[11] "WCT"         "WCT x COD"
```

We can reorder the levels by the following method. First, we load the *recode* function into R:

```
recode <- function(x,or){
   new.x <- factor(match(as.numeric(x),or))
   levels(new.x) <- levels(x)[or]
   new.x
}
```

Then we apply the *recode* function to the data:

```
checks <-
as.character(unique(augmented$treatment[augmented$type==”check”]))
or <- order(ifelse(levels(augmented$treatment)%in%checks,1,2))
augmented$treatment <- recode(augmented$treatment,or)
```

We can check that the levels are now reordered:

```
levels (augmented$treatment)
```

```
 [1] “COD x WCT”  “LCT x GBGD”  “WCT”  “WCT x COD”  “H1”
 [6] “H2”         “H3”          “H4”   “H5”         “H6”
[11] “H7”         “H8”
```

Perform the analysis and display the analysis of variance table

The first line defines the contrasts necessary to the analysis.

```
options(contrasts=c(“contr.helmert”,”contr.poly”))
augmented.aov <- aov(wax~block+treatment,data=augmented)
df.check <- length(checks)-1
df.treatment <- length(levels(augmented$treatment))-1
summary(augmented.aov,split=list(treatment=list(Check=1:df.check,”Test and test
vs. check”=(df.check+1):df.treatment)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	2	360.0714	180.0357	6.67486	0.029815*
Treatment	11	285.0952	25.9177	0.96091	0.549918
Treatment: Check	3	52.9167	17.6389	0.65396	0.609172
Treatment: Test and test vs. check	8	232.1786	29.0223	1.07601	0.477925
Residuals	6	161.8333	26.9722		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

The results are the same as in Table 8.7a. The p-values Pr(>F) are given directly, showing a block effect but no treatment effect.

Compute adjusted effects

Adjusted effects for both check treatments and test treatments are computed using the same commands as for BIBD.

```
options(contrasts=c("contr.sum","contr.poly"))
augmented.aov <- aov(wax~block+treatment,data=augmented)
co <- coef(augmented.aov)
co.treatment <- co[augmented.aov$assign==2]
effects.treatment <- c(co.treatment, -sum(co.treatment))
names(effects.treatment) <- levels(augmented$treatment)
print(data.frame(effects.treatment),digits=5)
```

	effects.treatment
COD x WCT	-2.0625
LCT x GBGD	0.9375
WCT	2.2708
WCT x COD	3.6042
H1	-2.8125
H2	5.4375
H3	-7.8125
H4	12.4375
H5	-3.8125
H6	-1.5625
H7	-2.8125
H8	-3.8125

The overall adjusted mean is in co[1]:

```
co[1]
```

```
(Intercept)
      81.0625
```

Adjusted block effects are computed using similar commands:

```
co.block <- co[augmented.aov$assign==1]
effects.block <- c(co.block, -sum(co.block))
names(effects.block) <- levels(augmented$block)
print(data.frame(effects.block),digits=5)
```

	effects.block
1	-3.25
2	0.75
3	2.50

Mean comparisons

We first compute the values necessary for comparing the means:

```
augmented.anova <- anova(augmented.aov)
sigma2 <- augmented.anova[[3]][3]
r <- augmented.anova$Df[1]+1 # Number of blocks
c <- length(checks) # Number of check treatments
t0 <- qt(0.975,augmented.aov$df.residual)
cd.test <- t0*sqrt(sigma2*(1+1/r+1/c+1/r/c))
cd.check <- t0*sqrt(2*sigma2/r)
print(c(sigma2,r,c,t0,cd.test,cd.check),digits=4)
```

```
[1] 26.972  3.000  4.000  2.447 16.406 10.376
```

Means have not to be compared since F tests are not significant. However, for demonstration, let us use the function *groups*

1) for check treatments

```
print(groups(co[1]+effects.treatment[1:(df.check+1)],cd.check),digits=5)
```

	x	groups
COD x WCT	79.000	a
LCT x GBGD	82.000	a
WCT	83.333	a
WCT x COD	84.667	a

2) for test treatements

```
print(groups(co[1]+effects.treatment[(df.check+2):(df.treatment+1)],cd.test),digits=5)
```

	x	groups
H3	73.25	a
H5	77.25	ab
H8	77.25	ab
H1	78.25	ab
H7	78.25	ab
H6	79.50	ab
H2	86.50	ab
H4	93.50	b

The 2 groups of test treatments are not to be considered since the F test is not significant.

Compute the anova of treatments ignoring blocks

This anova requires a particular set of contrasts. The following function, which has to be loaded into R, will be used to obtained the appropriate contrasts:

```
contr.augmented <- function(n1,n2){
  m1 <- contr.helmert(n1)
  m2 <- contr.helmert(n2)
  m10 <- cbind(m1,matrix(0,nrow(m1),ncol(m2)))
  m02 <- cbind(matrix(0,nrow(m2),ncol(m1)),m2)
  rbind(m10,m02)
}
```

The analysis can now be easily computed:

```
contrasts(augmented$treatment) <- contr.augmented(df.check+1,df.treatment-
df.check)
augmented.aov <- aov(wax~treatment+block,data=augmented)
summary(augmented.aov,split=list(treatment=list(Check=1:df.check,
Test=(df.check+1):(df.treatment-1),"Test vs. check"=df.treatment)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	11	575.6667	52.3333	1.94027	0.21468
Treatment: Check	3	52.9167	17.6389	0.65396	0.60917
Treatment: Test	7	505.8750	72.2679	2.67934	0.12526
Treatment: Test vs. check	1	16.8750	16.8750	0.62564	0.45907
Block	2	69.5000	34.7500	1.28836	0.34236
Residuals	6	161.8333	26.9722		

The results are the same as in Table 8.7b.

Contents of files used in the above computations and readable by R are printed below:

File 08-1-bibd.txt (data of Table 8.1)

block	treatment	seedlings
1	P7	40
1	P5	55
1	P4	65
2	P3	72
2	P5	58
2	P6	25
3	P5	63
3	P2	58
3	P1	67
4	P9	41
4	P3	80
4	P4	61
5	P5	52
5	P8	71
5	P9	49
6	P8	78
6	P7	46
6	P6	33
7	P8	69
7	P3	71
7	P2	61
8	P9	38
8	P1	70
8	P6	36
9	P9	34
9	P7	41
9	P2	52
10	P2	58
10	P4	68
10	P6	41
11	P3	74
11	P7	44
11	P1	71
12	P8	77
12	P1	61
12	P4	68

File 08-1-augmented.txt (data of Figure 8.1)

block	treatment	type	wax
1	"H8"	test	74
1	"LCT x GBGD"	check	78
1	"WCT"	check	78
1	"H3"	test	70
1	"WCT x COD"	check	83
1	"COD x WCT"	check	77
1	"H7"	test	75
2	"WCT"	check	91
2	"COD x WCT"	check	81
2	"WCT x COD"	check	79
2	"LCT x GBGD"	check	81
2	"H1"	test	79
2	"H5"	test	78
3	"H4"	test	96
3	"LCT x GBGD"	check	87
3	"WCT x COD"	check	92
3	"H2"	test	89
3	"WCT"	check	81
3	"COD x WCT"	check	79
3	"H6"	test	82

Appendix IX: Experimental designs for multiple factors

Factorial experiments

Load and display data

```
(factorial <- read.table("09-1-factorial.txt",header=T))
```

	block	irrigation	fertilizer	yield
1	B1	I1	F2	60
2	B1	I2	F0	55
3	B1	I0	F2	64
4	B1	I1	F0	63
5	B1	I0	F0	53
6	B1	I2	F2	71
7	B1	I1	F1	66
8	B1	I2	F1	65
9	B1	I0	F1	64
10	B2	I1	F1	58
11	B2	I1	F2	71
12	B2	I0	F2	58
13	B2	I2	F2	65
14	B2	I2	F0	53
15	B2	I0	F1	54
16	B2	I0	F0	52
17	B2	I2	F1	58
18	B2	I1	F0	59
19	B3	I2	F2	67
20	B3	I0	F2	61
21	B3	I2	F0	60
22	B3	I1	F1	67
23	B3	I1	F2	68
24	B3	I0	F0	45
25	B3	I2	F1	58
26	B3	I1	F0	63
27	B3	I0	F1	59

Perform the analysis and display the analysis of variance table

```
factorial.aov <- aov(yield~block+irrigation*fertilizer,data=factorial)
(factorial.anova <- anova(factorial.aov))
```

```
Analysis of Variance Table
Response: yield
                    Df  Sum Sq Mean Sq F value    Pr(>F)
Block                2  61.4074 30.7037 1.99759 0.16809593
Irrigation           2 241.4074 120.7037 7.85301 0.00420558 **
Fertilizer           2 375.4074 187.7037 12.21205 0.00060233 ***
Irrigation:fertilizer 4  72.1481 18.0370 1.17349 0.35944206
Residuals           16 245.9259 15.3704

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 9.2. The main effects of Irrigation and Fertilizer are significantly different. The interaction between these two factors is not significant.

Compute the means

```
options(digits=4)
model.tables(factorial.aov,type="means")$tables
```

```
$"Grand mean"
[1] 60.63

$block
block
   B1    B2    B3
62.33 58.67 60.89

$irrigation
irrigation
   I0    I1    I2
56.67 63.89 61.33

$fertilizer
fertilizer
   F0    F1    F2
55.89 61.00 65.00

$"irrigation:fertilizer"
          fertilizer
irrigation    F0    F1    F2
        I0 50.00 59.00 61.00
        I1 61.67 63.67 66.33
        I2 56.00 60.33 67.67
```

The means are computed for the blocks and for the different levels of irrigation, fertilizer, and irrigation:fertilizer. These results correspond to Table 9.1, in which only the totals are computed.

Compute the CD (5%) for comparing Irrigation levels

Some values stored in *factorial.anova* can be accessed by their position in the model (1 for block, 3 for fertilizer, and 5 for error)

```
b <- factorial.anova$Df[1]+1 # Number of blocks
f <- factorial.anova$Df[3]+1 # Number of fertilizer levels
sigma2 <- factorial.anova$"Mean Sq"[5] # Error mean square
t0 <- qt(0.975,factorial.aov$df.residual)
cd <- t0*sqrt(2*sigma2/b/f)
c(t0,sigma2,b,f,cd)
```

```
[1]  2.120 15.370  3.000  3.000  3.918
```

Perform the analysis and display the analysis of variance table with linear (L) and quadratic (Q) contrasts

As the levels of irrigation and fertilizer are quantitative, their effects can be interpreted with contrasts. For example, consider the set of linear and quadratic contrasts (polynomial contrasts).

```
options(digits=9)
contrasts(factorial$irrigation) <- contr.poly(3)
contrasts(factorial$fertilizer) <- contr.poly(3)
factorial.aov <- aov(yield~block+irrigation*fertilizer,data=factorial)
summary(factorial.aov,split=list(irrigation=list(L=1,Q=2),
fertilizer=list(L=1,Q=2)),expand.split=F)
```

```
                    Df    Sum Sq   Mean Sq  F value     Pr(>F)
Block                2   61.4074   30.7037  1.99759 0.16809593
Irrigation           2  241.4074  120.7037  7.85301 0.00420558 **
  Irrigation: L      1   98.0000   98.0000  6.37590 0.02250901 *
  Irrigation: Q      1  143.4074  143.4074  9.33012 0.00756683 **
Fertilizer           2  375.4074  187.7037 12.21205 0.00060233 ***
  Fertilizer: L      1  373.5556  373.5556 24.30361 0.00015079 ***
  Fertilizer: Q      1    1.8519    1.8519  0.12048 0.73303591
Irrigation:Fertilizer 4  72.1481   18.0370  1.17349 0.35944206
Residuals           16  245.9259   15.3704

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The 2 degrees of freedom of irrigation and fertilizer are split into a linear and a quadratic component, both with 1 df. For irrigation, both linear and quadratic effects are significant, indicating significant differences I2-I0 and (I0+I2)/2-I1. For fertilizer, only the linear effect is significant, indicating a purely linear response to fertilizer.

Split Plot Design

Load and display data

```
(splitplot <- read.table("09-3-splitplot.txt",header=T))
```

	main	block	irrigation	fertilizer	yield
1	P1	B1	I0	F0	15.9
2	P1	B1	I0	F1	21.1
3	P1	B1	I0	F2	18.0
4	P2	B1	I1	F0	14.8
5	P2	B1	I1	F1	19.3
6	P2	B1	I1	F2	17.3
7	P3	B1	I2	F0	8.1
8	P3	B1	I2	F1	15.1
9	P3	B1	I2	F2	15.8
10	P4	B2	I0	F0	15.2
11	P4	B2	I0	F1	20.0
12	P4	B2	I0	F2	19.7
13	P5	B2	I1	F0	14.0
14	P5	B2	I1	F1	18.6
15	P5	B2	I1	F2	15.8
16	P6	B2	I2	F0	7.2
17	P6	B2	I2	F1	12.7
18	P6	B2	I2	F2	12.3
19	P7	B3	I0	F0	13.8
20	P7	B3	I0	F1	19.2
21	P7	B3	I0	F2	17.1
22	P8	B3	I1	F0	15.0
23	P8	B3	I1	F1	18.2
24	P8	B3	I1	F2	18.5
25	P9	B3	I2	F0	9.4
26	P9	B3	I2	F1	14.4
27	P9	B3	I2	F2	16.0

Note that a column *main* has been included in the data in order to identify the main plots.

Perform the analysis and display the analysis of variance table

```
splitplot.aov <-
aov(yield~block+fertilizer*irrigation+Error(main),data=splitplot)summary(splitplot.aov)
```

```
Error: main
                    Df    Sum Sq   Mean Sq  F value    Pr(>F)
Block               2   5.54296   2.77148  0.81920  0.503278
Irrigation          2 152.35185  76.17593 22.51628  0.006655 **
Residuals           4  13.53259   3.38315

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
                    Df    Sum Sq   Mean Sq  F value    Pr(>F)
Fertilizer          2 129.07630  64.53815 103.10828 2.7654e-08 ***
Fertilizer:Irrigation 4  13.59259   3.39815   5.42899  0.009888 **
Residuals           12   7.51111   0.62593

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results are the same as in Table 9.5. The main effects of Irrigation and Fertilizer are significantly different. The interaction between these two factors is also significant.

Compute the means

```
options(digits=4)
model.tables(splitplot.aov,type="means")$tables
```

```
$"Grand mean"
[1] 15.65

$block
block
    B1    B2    B3
 16.16 15.06 15.73
$fertilizer
fertilizer
    F0    F1    F2
 12.60 17.62 16.72
$irrigation
irrigation
    I0    I1    I2
 17.78 16.83 12.33
```

```
$"fertilizer:irrigation"
            irrigation
fertilizer          I0          I1          I2
        F0      14.967      14.600       8.233
        F1      20.100      18.700      14.067
        F2      18.267      17.200      14.700
```

The means are computed for the blocks and for the different levels of irrigation, fertilizer, and irrigation:fertilizer. These results correspond to Table 9.6.

In order to compute the CDs (5%), some values have to be extracted from main plot and subplot anovas.

Extract values from main plot anova

```
splitplot.anova.main <- unlist(summary(splitplot.aov$main))
r <- splitplot.anova.main["Df1"]+1 # Number of blocks
a <- splitplot.anova.main["Df2"]+1 # Number of irrigation levels
df.main <- splitplot.anova.main["Df3"] # df of main plot error
sigma2.main <- splitplot.anova.main["Mean Sq3"] # main plot error
t1 <- qt(0.975,df.main)
as.numeric(c(r,a,df.main,sigma2.main,t1))
```

```
[1] 3.000 3.000 4.000 3.383 2.776
```

Extract values from subplot anova

```
splitplot.anova.subplot <- unlist(summary(splitplot.aov$Within))
b <- splitplot.anova.subplot["Df1"]+1 # Number of fertilizer levels
df.subplot <- splitplot.anova.subplot["Df3"] # df of subplot error
sigma2.subplot <- splitplot.anova.subplot["Mean Sq3"] # subplot error
t2 <- qt(0.975,df.subplot)
as.numeric(c(b,df.subplot,sigma2.subplot,t2))
```

```
[1]  3.0000 12.0000  0.6259  2.1788
```

Compute the CDs

```
cd.main  <-  t1*sqrt(2*sigma2.main/b/r)
cd.subplot  <-  t2*sqrt(2*sigma2.subplot/a/r)
t <- ((b-1)*t2*sigma2.subplot+t1*sigma2.main)/((b-1)*sigma2.subplot +sigma2.main)
cd.main.subplot  <-  t*sqrt(2*((b-1)*sigma2.subplot+sigma2.main)/b/r)
cd.subplot.main  <-  t2*sqrt(2*sigma2.subplot/r)
as.numeric(c(cd.main,cd.subplot,cd.main.subplot,cd.subplot.main))
```

```
[1]  2.4074  0.8126  2.6540  1.4075
```

Strip Plot Design

Load and display data

```
(stripplot  <-  read.table("09-7-stripplot.txt",header=T))
```

	main1	main2	fertilizer	cultivar	irrigation	block	nuts
1	P1	Q1	F1	C1	R	B1	77
2	P1	Q2	F1	C1	I	B1	91
3	P2	Q1	F1	C2	R	B1	44
4	P2	Q2	F1	C2	I	B1	115
5	P3	Q1	F1	C3	R	B1	64
6	P3	Q2	F1	C3	I	B1	76
7	P4	Q1	F2	C1	R	B1	111
8	P4	Q2	F2	C1	I	B1	115
9	P5	Q1	F2	C2	R	B1	139
10	P5	Q2	F2	C2	I	B1	171
11	P6	Q1	F2	C3	R	B1	100
12	P6	Q2	F2	C3	I	B1	116
13	P7	Q1	F3	C1	R	B1	131
14	P7	Q2	F3	C1	I	B1	133
15	P8	Q1	F3	C2	R	B1	178
16	P8	Q2	F3	C2	I	B1	154
17	P9	Q1	F3	C3	R	B1	147
18	P9	Q2	F3	C3	I	B1	133
19	P10	Q3	F1	C1	R	B2	63
20	P10	Q4	F1	C1	I	B2	114

21	P11	Q3	F1	C2	R	B2	92
22	P11	Q4	F1	C2	I	B2	109
23	P12	Q3	F1	C3	R	B2	61
24	P12	Q4	F1	C3	I	B2	113
25	P13	Q3	F2	C1	R	B2	98
26	P13	Q4	F2	C1	I	B2	153
27	P14	Q3	F2	C2	R	B2	113
28	P14	Q4	F2	C2	I	B2	123
29	P15	Q3	F2	C3	R	B2	118
30	P15	Q4	F2	C3	I	B2	141
31	P16	Q3	F3	C1	R	B2	123
32	P16	Q4	F3	C1	I	B2	132
33	P17	Q3	F3	C2	R	B2	114
34	P17	Q4	F3	C2	I	B2	171
35	P18	Q3	F3	C3	R	B2	101
36	P18	Q4	F3	C3	I	B2	177
37	P19	Q5	F1	C1	R	B3	132
38	P19	Q6	F1	C1	I	B3	128
39	P20	Q5	F1	C2	R	B3	92
40	P20	Q6	F1	C2	I	B3	122
41	P21	Q5	F1	C3	R	B3	105
42	P21	Q6	F1	C3	I	B3	141
43	P22	Q5	F2	C1	R	B3	118
44	P22	Q6	F2	C1	I	B3	119
45	P23	Q5	F2	C2	R	B3	167
46	P23	Q6	F2	C2	I	B3	151
47	P24	Q5	F2	C3	R	B3	138
48	P24	Q6	F2	C3	I	B3	134
49	P25	Q5	F3	C1	R	B3	133
50	P25	Q6	F3	C1	I	B3	162
51	P26	Q5	F3	C2	R	B3	144
52	P26	Q6	F3	C2	I	B3	167
53	P27	Q5	F3	C3	R	B3	93
54	P27	Q6	F3	C3	I	B3	145

Note that two columns have been included in the data. The column *main1* identifies the main plots with same fertilizer x cultivar combinations within blocks; the column *main2* identifies the main plots with same irrigation levels within blocks.

The analysis of variance is computed in 3 steps.

Step 1. Analysis of variance of treatment combinations

```
options(digits=9)
stripplot.aov1 <-
aov(nuts~block+fertilizer:cultivar+Error(main1),data=stripplot)
summary(stripplot.aov1)[[1]]
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	2	3031.148	1515.574	3.06453	0.0746888 .
Fertilizer:Cultivar	8	22977.370	2872.171	5.80761	0.0014138 **
Residuals	16	7912.852	494.553		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step 2. Analysis of variance of irrigation

```
stripplot.aov2 <- aov(nuts~block+irrigation+Error(main2),data=stripplot)
anova(stripplot.aov2[[2]])
```

Analysis of Variance Table

Response: nuts

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Irrigation	1	6890.741	6890.741	7.55272	0.11082
Residuals	2	1824.704	912.352		

Step 3. Analysis of variance of treatment by irrigation

```
stripplot.aov3 <- aov(nuts~block+irrigation+fertilizer:cultivar +irrigation:fertilizer:
cultivar+block:irrigation+block:fertilizer:cultivar,data=stripplot)
anova(stripplot.aov3)[c(5,7),]
```

Analysis of Variance Table

Response: nuts

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Irrigation:Fertilizer:Cultivar	8	1588.926	198.616	0.54058	0.80947
Residuals	16	5878.630	367.414		

The results of the 3 steps are the same as in Table 9.8.

Compute the means

```
options(digits=4)
model.tables(stripplot.aov3,type="means")$tables[1:4]
```

```
$"Grand mean"
[1] 122.3

$block
block
   B1    B2    B3
116.4 117.6 132.8

$irrigation
irrigation
    I     R
133.6 111.0

$"fertilizer:cultivar"
          cultivar
fertilizer    C1     C2     C3
        F1 100.83  95.67  93.33
        F2 119.00 144.00 124.50
        F3 135.67 154.67 132.67
```

Compute the CD (5%) for treatment combinations

```
stripplot.anova1 <- unlist(summary(stripplot.aov1)[[1]])
df.main1 <- stripplot.anova1["Df3"] # df of main1 plot error
sigma2.main1 <- stripplot.anova1["Mean Sq3"] # main1 plot error
t1 <- qt(0.975,df.main1)
stripplot.anova2 <- unlist(summary(stripplot.aov2)[[1]])
r <- stripplot.anova2["Df1"]+1 # Number of blocks
b <- stripplot.anova2["Df2"]+1 # Number of levels of irrigation
cd.main1 <- t1*sqrt(2*sigma2.main1/r/b)
as.numeric(c(r,b,df.main1,sigma2.main1,t1,cd.main1))
```

```
[1]   3.00   2.00  16.00 494.55   2.12  27.22
```

Contents of files used in the above computations and readable by R are printed below:

File 09-1-factorial.txt (data of Fig. 9.1)

block	irrigation	fertilizer	yield
B1	I1	F2	60
B1	I2	F0	55
B1	I0	F2	64
B1	I1	F0	63
B1	I0	F0	53
B1	I2	F2	71
B1	I1	F1	66
B1	I2	F1	65
B1	I0	F1	64
B2	I1	F1	58
B2	I1	F2	71
B2	I0	F2	58
B2	I2	F2	65
B2	I2	F0	53
B2	I0	F1	54
B2	I0	F0	52
B2	I2	F1	58
B2	I1	F0	59
B3	I2	F2	67
B3	I0	F2	61
B3	I2	F0	60
B3	I1	F1	67
B3	I1	F2	68
B3	I0	F0	45
B3	I2	F1	58
B3	I1	F0	63
B3	I0	F1	59

File 09-3-splitplot.txt (data of Table 9.3)

main	block	irrigation	fertilizer	yield
P1	B1	I0	F0	15.9
P1	B1	I0	F1	21.1
P1	B1	I0	F2	18.0
P2	B1	I1	F0	14.8
P2	B1	I1	F1	19.3
P2	B1	I1	F2	17.3
P3	B1	I2	F0	8.1
P3	B1	I2	F1	15.1
P3	B1	I2	F2	15.8
P4	B2	I0	F0	15.2
P4	B2	I0	F1	20.0
P4	B2	I0	F2	19.7
P5	B2	I1	F0	14.0
P5	B2	I1	F1	18.6
P5	B2	I1	F2	15.8
P6	B2	I2	F0	7.2
P6	B2	I2	F1	12.7
P6	B2	I2	F2	12.3
P7	B3	I0	F0	13.8
P7	B3	I0	F1	19.2
P7	B3	I0	F2	17.1
P8	B3	I1	F0	15.0
P8	B3	I1	F1	18.2
P8	B3	I1	F2	18.5
P9	B3	I2	F0	9.4
P9	B3	I2	F1	14.4
P9	B3	I2	F2	16.0

File 09-7-stripplot.txt (data of Table 9.3)

main1	main2	fertilizer	cultivar	irrigation	block	nuts
P1	Q1	F1	C1	R	B1	77
P1	Q2	F1	C1	I	B1	91
P2	Q1	F1	C2	R	B1	44
P2	Q2	F1	C2	I	B1	115
P3	Q1	F1	C3	R	B1	64
P3	Q2	F1	C3	I	B1	76
P4	Q1	F2	C1	R	B1	111
P4	Q2	F2	C1	I	B1	115
P5	Q1	F2	C2	R	B1	139
P5	Q2	F2	C2	I	B1	171

P6	Q1	F2	C3	R	B1	100
P6	Q2	F2	C3	I	B1	116
P7	Q1	F3	C1	R	B1	131
P7	Q2	F3	C1	I	B1	133
P8	Q1	F3	C2	R	B1	178
P8	Q2	F3	C2	I	B1	154
P9	Q1	F3	C3	R	B1	147
P9	Q2	F3	C3	I	B1	133
P10	Q3	F1	C1	R	B2	63
P10	Q4	F1	C1	I	B2	114
P11	Q3	F1	C2	R	B2	92
P11	Q4	F1	C2	I	B2	109
P12	Q3	F1	C3	R	B2	61
P12	Q4	F1	C3	I	B2	113
P13	Q3	F2	C1	R	B2	98
P13	Q4	F2	C1	I	B2	153
P14	Q3	F2	C2	R	B2	113
P14	Q4	F2	C2	I	B2	123
P15	Q3	F2	C3	R	B2	118
P15	Q4	F2	C3	I	B2	141
P16	Q3	F3	C1	R	B2	123
P16	Q4	F3	C1	I	B2	132
P17	Q3	F3	C2	R	B2	114
P17	Q4	F3	C2	I	B2	171
P18	Q3	F3	C3	R	B2	101
P18	Q4	F3	C3	I	B2	177
P19	Q5	F1	C1	R	B3	132
P19	Q6	F1	C1	I	B3	128
P20	Q5	F1	C2	R	B3	92
P20	Q6	F1	C2	I	B3	122
P21	Q5	F1	C3	R	B3	105
P21	Q6	F1	C3	I	B3	141
P22	Q5	F2	C1	R	B3	118
P22	Q6	F2	C1	I	B3	119
P23	Q5	F2	C2	R	B3	167
P23	Q6	F2	C2	I	B3	151
P24	Q5	F2	C3	R	B3	138
P24	Q6	F2	C3	I	B3	134
P25	Q5	F3	C1	R	B3	133
P25	Q6	F3	C1	I	B3	162
P26	Q5	F3	C2	R	B3	144
P26	Q6	F3	C2	I	B3	167
P27	Q5	F3	C3	R	B3	93
P27	Q6	F3	C3	I	B3	145

Appendix X: Analysis of multilocation trials

Read and display data

```
(multiloc <- read.table("10-4-multiloc.txt",header=T))
```

	location	genotype	rep	yield
1	Location-1	LCT	rep-1	40.6
2	Location-1	LCT	rep-2	77.0
3	Location-1	LCT	rep-3	24.5
4	Location-1	LCT	rep-4	52.5
5	Location-1	CCT	rep-1	58.4
6	Location-1	CCT	rep-2	75.7
.				
.				
.				
91	Location-4	FJT	rep-3	76.2
92	Location-4	FJT	rep-4	84.4
93	Location-4	JMT	rep-1	102.6
94	Location-4	JMT	rep-2	110.3
95	Location-4	JMT	rep-3	108.8
96	Location-4	JMT	rep-4	92.1

(See complete data below)

Analysis of variance

```
options(digits=9)
multiloc.aov <- aov(yield~genotype*location+rep%in%location,data=multiloc)
anova(multiloc.aov)
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genotype	5	10201.14	2040.23	9.81819	6.8345e-07 ***
Location	3	36220.33	12073.44	58.10105	< 2.22e-16 ***
Genotype:Location	15	17279.46	1151.96	5.54360	7.2992e-07 ***
Location:rep	12	7814.59	651.22	3.13385	0.0016598 **
Residuals	60	12468.05	207.80		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results are the same as in Table 10.5.

Estimates of regression coefficients and corresponding regression SS

The first step is to compute R objects containing m_i, m_j, and m_{ij}

```
df.mij <-aggregate(multiloc$yield,by=list(location=multiloc$location,
          genotype=multiloc$genotype),FUN=mean)
df.mi <- aggregate(multiloc$yield,by=list(genotype=multiloc$genotype),
          FUN=mean)
df.mj <- aggregate(multiloc$yield,by=list(location=multiloc$location),
          FUN=mean)
df.mj.mij <- merge(df.mij,df.mj,by="location")
names(df.mj.mij)[3:4] <- c("mij","mj")
df.mi.mj.mij <- merge(df.mj.mij,df.mi,by="genotype")
names(df.mi.mj.mij)[5] <- "mi"
multiloc.mi.mj.mij <- merge(multiloc,df.mi.mj.mij,
                            by=c("genotype","location"))
multiloc.mi.mj.mij
```

	genotype	location	rep	yield	mij	mj	mi
1	AOT	Location-1	rep-1	78.0	80.800	70.7166667	104.80625
2	AOT	Location-1	rep-2	91.8	80.800	70.7166667	104.80625
3	AOT	Location-1	rep-3	81.4	80.800	70.7166667	104.80625
4	AOT	Location-1	rep-4	72.0	80.800	70.7166667	104.80625
5	AOT	Location-2	rep-1	48.0	71.325	71.3625000	104.80625
6	AOT	Location-2	rep-2	88.3	71.325	71.3625000	104.80625
.							
.							
.							
91	PHOT	Location-3	rep-3	105.0	113.725	116.6125000	96.04375
92	PHOT	Location-3	rep-4	121.2	113.725	116.6125000	96.04375
93	PHOT	Location-4	rep-1	100.7	96.400	99.1083333	96.04375
94	PHOT	Location-4	rep-2	81.2	96.400	99.1083333	96.04375
95	PHOT	Location-4	rep-3	123.5	96.400	99.1083333	96.04375
96	PHOT	Location-4	rep-4	80.2	96.400	99.1083333	96.04375

The regressions will be computed with the following function, which has to be loaded into R

```
reg <- function(df){
  res.lm <- lm(mij~mj,data=df)
  res.anova <- anova(res.lm)
  res <- c(res.lm$coeff[[2]],res.anova[[2]],res.anova[[5]][1])
  names(res) <- c("Regression coeff.","Regression SS","Residual SS","Pr(>F)")
  res
}
```

We can now compute the regressions

```
df.mj.mij.split <- split(df.mj.mij,df.mj.mij$genotype)
round(x <- sapply(df.mj.mij.split,reg),3)
```

	AOT	CCT	FJT	JMT	LCT	PHOT
Regression coeff.	1.512	1.126	0.240	1.592	0.993	0.537
Regression SS	3451.035	1914.840	86.597	3824.349	1488.154	435.179
Residual SS	83.946	25.787	380.874	442.633	1062.689	178.862
Pr(>F)	0.012	0.007	0.570	0.053	0.236	0.158

These results are the same as in Table 10.6.

Anova for regression analysis

```
res <- lm(yield~genotype+location+mi*mj+location:mi+genotype:mj
          +genotype:location+rep%in%location,data=multiloc.mi.mj.mij)
anova(res)
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genotype	5	10201.14	2040.23	9.81819	6.8345e-07 ***
Location	3	36220.33	12073.44	58.10105	< 2.22e-16 ***
mi:mj	1	327.16	327.16	1.57438	0.21443782
Location:mi	2	950.58	475.29	2.28724	0.11033440
Genotype:mj	4	8253.13	2063.28	9.92914	3.1171e-06 ***
Genotype:location	8	7748.59	968.57	4.66107	0.00018159 ***
Location:rep	12	7814.59	651.22	3.13385	0.00165977 **
Residuals	60	12468.05	207.80		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Contents of files used in the above computations and readable by R are printed below:

File 10-4-multiloc.txt (data of Table 10.4)

location	genotype	rep	yield
Location-1	LCT	rep-1	40.6
Location-1	LCT	rep-2	77.0
Location-1	LCT	rep-3	24.5
Location-1	LCT	rep-4	52.5

Location-1	CCT	rep-1	58.4
Location-1	CCT	rep-2	75.7
Location-1	CCT	rep-3	58.0
Location-1	CCT	rep-4	86.5
Location-1	AOT	rep-1	78.0
Location-1	AOT	rep-2	91.8
Location-1	AOT	rep-3	81.4
Location-1	AOT	rep-4	72.0
Location-1	PHOT	rep-1	90.0
Location-1	PHOT	rep-2	105.0
Location-1	PHOT	rep-3	86.9
Location-1	PHOT	rep-4	99.5
Location-1	FJT	rep-1	92.0
Location-1	FJT	rep-2	123.9
Location-1	FJT	rep-3	61.1
Location-1	FJT	rep-4	97.2
Location-1	JMT	rep-1	25.9
Location-1	JMT	rep-2	50.4
Location-1	JMT	rep-3	46.9
Location-1	JMT	rep-4	22.0
Location-2	LCT	rep-1	97.9
Location-2	LCT	rep-2	114.0
Location-2	LCT	rep-3	94.4
Location-2	LCT	rep-4	74.4
Location-2	CCT	rep-1	76.0
Location-2	CCT	rep-2	75.0
Location-2	CCT	rep-3	43.2
Location-2	CCT	rep-4	58.7
Location-2	AOT	rep-1	48.0
Location-2	AOT	rep-2	88.3
Location-2	AOT	rep-3	79.8
Location-2	AOT	rep-4	69.2
Location-2	PHOT	rep-1	94.0
Location-2	PHOT	rep-2	55.6
Location-2	PHOT	rep-3	87.3
Location-2	PHOT	rep-4	77.9
Location-2	FJT	rep-1	80.9
Location-2	FJT	rep-2	77.9
Location-2	FJT	rep-3	88.0
Location-2	FJT	rep-4	73.0
Location-2	JMT	rep-1	44.7
Location-2	JMT	rep-2	28.3
Location-2	JMT	rep-3	47.0
Location-2	JMT	rep-4	39.2
Location-3	LCT	rep-1	107.0
Location-3	LCT	rep-2	132.0

Location-3	LCT	rep-3	117.0
Location-3	LCT	rep-4	115.3
Location-3	CCT	rep-1	122.0
Location-3	CCT	rep-2	129.0
Location-3	CCT	rep-3	111.0
Location-3	CCT	rep-4	110.0
Location-3	AOT	rep-1	94.0
Location-3	AOT	rep-2	187.0
Location-3	AOT	rep-3	153.0
Location-3	AOT	rep-4	138.7
Location-3	PHOT	rep-1	96.7
Location-3	PHOT	rep-2	132.0
Location-3	PHOT	rep-3	105.0
Location-3	PHOT	rep-4	121.2
Location-3	FJT	rep-1	98.2
Location-3	FJT	rep-2	106.0
Location-3	FJT	rep-3	115.0
Location-3	FJT	rep-4	95.8
Location-3	JMT	rep-1	73.0
Location-3	JMT	rep-2	97.0
Location-3	JMT	rep-3	124.0
Location-3	JMT	rep-4	118.8
Location-4	LCT	rep-1	97.8
Location-4	LCT	rep-2	97.0
Location-4	LCT	rep-3	96.9
Location-4	LCT	rep-4	93.9
Location-4	CCT	rep-1	114.0
Location-4	CCT	rep-2	93.2
Location-4	CCT	rep-3	102.8
Location-4	CCT	rep-4	80.6
Location-4	AOT	rep-1	129.3
Location-4	AOT	rep-2	132.7
Location-4	AOT	rep-3	143.7
Location-4	AOT	rep-4	90.0
Location-4	PHOT	rep-1	100.7
Location-4	PHOT	rep-2	81.2
Location-4	PHOT	rep-3	123.5
Location-4	PHOT	rep-4	80.2
Location-4	FJT	rep-1	82.7
Location-4	FJT	rep-2	64.0
Location-4	FJT	rep-3	76.2
Location-4	FJT	rep-4	84.4
Location-4	JMT	rep-1	102.6
Location-4	JMT	rep-2	110.3
Location-4	JMT	rep-3	108.8
Location-4	JMT	rep-4	92.1

Appendix XI: Multivariate analysis and determination of genetic distance

Load data

```
coconuts <- read.table("11-1-coconuts.txt",header=T)
variables <- as.matrix(coconuts[,-1])
accession <- coconuts$accession
```

Display data

coconuts

	accession	fruit.weight	fruit.length	husk.thickness	husk.weight
1	SSAT	984.50	26.875	2.250	245.8
2	SSAT	1040.00	32.500	2.725	329.0
3	SSAT	712.00	25.875	2.250	192.8
4	SSAT	1100.25	28.750	2.475	280.0
5	POLT	765.25	27.875	3.650	270.0
6	POLT	713.67	29.500	3.900	333.3
7	POLT	669.50	28.125	3.425	271.5
8	POLT	629.50	29.250	3.600	272.5
9	MVT	1591.67	33.833	3.167	373.3
10	MVT	1589.25	32.250	2.925	384.3
11	MVT	2372.50	35.250	4.225	893.5
12	MVT	1723.25	33.500	3.300	502.0
13	KKT	1407.50	32.250	2.600	401.8
14	KKT	1863.75	34.125	3.550	615.0
15	KKT	1069.50	29.375	2.875	328.3
16	KKT	1395.50	31.500	3.225	475.5
17	NLAD	980.75	30.500	3.225	413.3
18	NLAD	963.50	31.250	3.062	432.5
19	NLAD	1047.25	31.500	2.925	410.8
20	NLAD	1056.50	31.375	3.362	472.3

variables

	fruit.weight	fruit.length	husk.thickness	husk.weight
1	984.50	26.875	2.250	245.8
2	1040.00	32.500	2.725	329.0
3	712.00	25.875	2.250	192.8
4	1100.25	28.750	2.475	280.0
5	765.25	27.875	3.650	270.0

6	713.67	29.500	3.900	333.3
7	669.50	28.125	3.425	271.5
8	629.50	29.250	3.600	272.5
9	1591.67	33.833	3.167	373.3
10	1589.25	32.250	2.925	384.3
11	2372.50	35.250	4.225	893.5
12	1723.25	33.500	3.300	502.0
13	1407.50	32.250	2.600	401.8
14	1863.75	34.125	3.550	615.0
15	1069.50	29.375	2.875	328.3
16	1395.50	31.500	3.225	475.5
17	980.75	30.500	3.225	413.3
18	963.50	31.250	3.062	432.5
19	1047.25	31.500	2.925	410.8
20	1056.50	31.375	3.362	472.3

```
accession
```

```
[1] SSAT SSAT SSAT SSAT POLT POLT POLT POLT MVT MVT MVT MVT KKT KKT KKT
[16] KKT  NLAD NLAD NLAD NLAD
Levels: KKT MVT NLAD POLT SSAT
```

Compute manova and Wilks' statistic (with F test)

```
options(digits=7)
fit <- manova(variables~accession)
(result <- summary(fit,test="Wilks"))
```

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
Accession	4.000	0.0052	10.7164	16.000	37.298	1.813e-09***
Residuals	15.000					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The value of Wilks' statistic (0.0052) differs slightly from the value obtained previously in this chapter (0.00565) because the SSSP matrices had only one decimal digit, but has 4 decimal digits in R. It is easy to check that if we keep only one decimal digit in R the results are the same. The slight differences observed in the other results have the same origin. Here the Wilks' statistic is tested with an F test, which provides a better approximation than chi square (Rao 1951). However it is possible to compute the chi square test (see below).

Compute D = Between Accessions SSSP matrix and W = Error SSSP matrix

```
(D <- result$SS$accession)
```

	fruit.weight	fruit.length	husk.thickness	husk.weight
fruit.weight	3142893.0177	14361.156142	248.007100	730073.6595
fruit.length	14361.1561	77.435890	5.084169	4101.9022
husk.thickness	248.0071	5.084169	3.362990	295.9992
husk.weight	730073.6595	4101.902200	295.999200	219795.2450

```
(W <- result$SS$Residuals)
```

	fruit.weight	fruit.length	husk.thickness	husk.weight
fruit.weight	844670.3352	4002.328957	999.374667	415802.411
fruit.length	4002.3290	44.292886	5.456996	1954.641
husk.thickness	999.3747	5.456996	1.861579	601.559
husk.weight	415802.4113	1954.641450	601.559000	238499.392

D and W are equal (with more digits) to the corresponding matrices in Table 11.2. The chi square test of Wilks' statistic is not necessary since we already have the F test, but it can be computed as follows, with the same symbols as previously in this chapter.

Compute chi square test for Wilks' statistic

```
lambda <- det(W)/det(D+W)
n <- nrow(variables)
p <- ncol(variables)
k <- fit$rank
m <- n-1-(p+k)/2
v <- -m*log(lambda)
df <- p*(k-1)
c(n,p,k,m,v,df,qchisq(0.95,df))
```

```
[1] 20.00000  4.00000  5.00000 14.50000 76.29241 16.00000 26.29623
```

We reject the null hypothesis since 26.29623 < 76.29241 (26.296 < 75.053 previously).

Compute X = matrix of average values

```
X <- aggregate(variables,by=list(accession),FUN=mean)
rownames(X) <- X[,1]
X <- X[-1]
X
```

	fruit.weight	fruit.length	husk.thickness	husk.weight
KKT	1434.0625	31.81250	3.06250	455.150
MVT	1819.1675	33.70825	3.40425	538.275
NLAD	1012.0000	31.15625	3.14350	432.225
POLT	694.4800	28.68750	3.64375	286.825
SSAT	959.1875	28.50000	2.42500	261.900

The values are identical to the values in Table 11.4.

Compute D2 = Mahalanobis' generalized distance matrix

```
S <- W/fit$df.residual
D2 <- as.dist(apply(X,1,function(u) mahalanobis(X,u,S)))
D2
```

	KKT	MVT	NLAD	POLT
MVT	10.313171			
NLAD	19.722993	54.663303		
POLT	59.403815	79.719746	40.170966	
SSAT	6.909978	31.603402	13.546829	61.509788

The distances are identical to those in Table 11.5 (but the orders or rows and columns are different).

Compute dendrogram

As R does not divide the distance by 2 we have to do this in line 2 in order to obtain the same result as previously

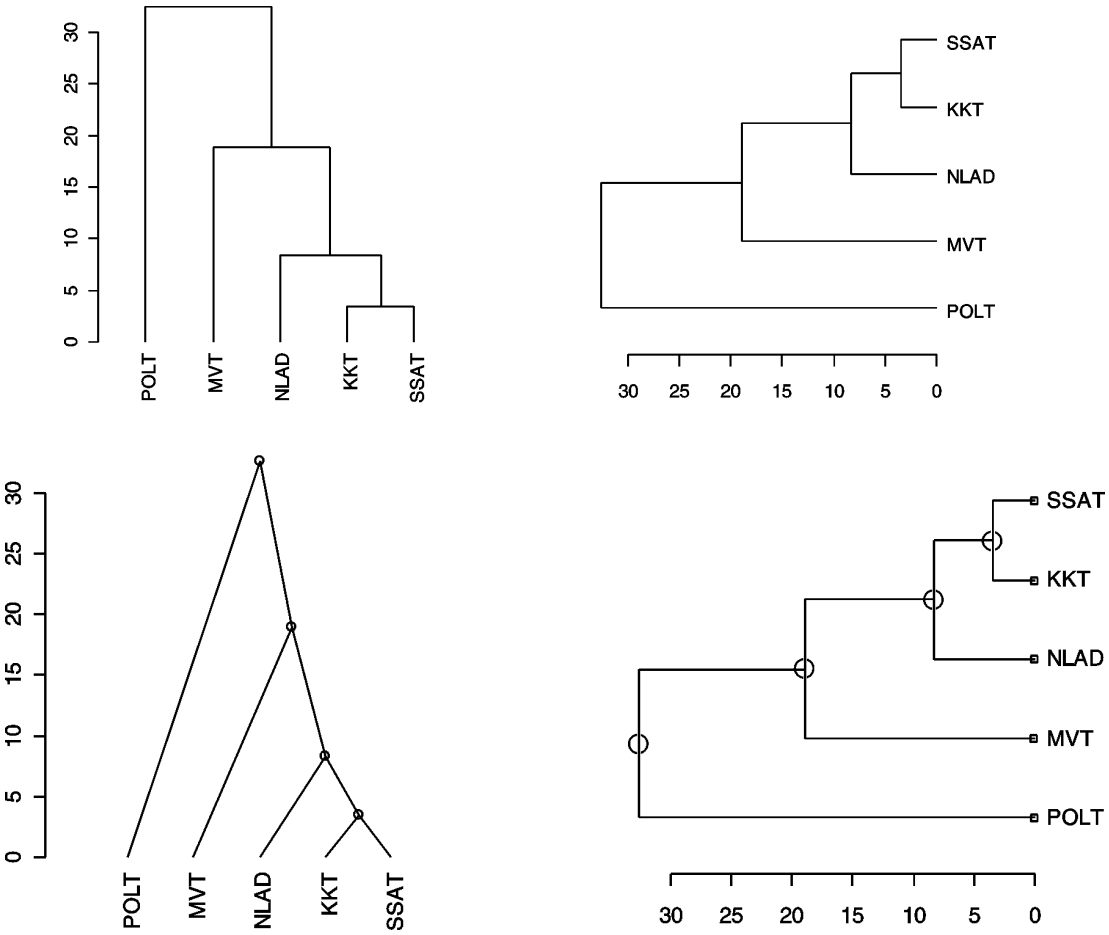
```
hc <- hclust(D2,method="mcquitty")
hc$height <- hc$height/2
dend <- as.dendrogram(hc)
str(dend)
```

```
—[dendrogram w/2 branches and 5 members at h = 32.5]
  |—leaf "POLT"
  '—[dendrogram w/2 branches and 4 members at h = 18.9]
      |—leaf "MVT"
      '—[dendrogram w/2 branches and 3 members at h = 8.32]
          |—leaf "NLAD"
          '—[dendrogram w/2 branches and 2 members at h = 3.45]
              |—leaf "KKT"
              '—leaf "SSAT"
```

This output corresponds to Table 11.7. The distance between POLT and group C-3 is 32.5 instead of 31.57 previously. The reason is that with the McQuitty method used in R, the distance between a leaf and a complex group is half the sum of the distances between the leaf and the 2 sub-groups of the group.

Plot dendrogram (examples of 4 methods)

```
op <- par(mfrow=c(2,2))
plot(dend)
plot(dend,horiz=T)
plot(dend,nodePar=list(pch=c(1,NA)),type="t",center=TRUE)
nP <- list(col=3:2,cex=c(2.0,0.75),pch= 21:22,bg=c("light blue","pink"),
           lab.col="tomato")
plot(dend,nodePar=nP,edgePar=list(col="gray",lwd=2),horiz=T)
par(op)
```



Contents of files used in the above computations and readable by R are printed below:

File 11-1-coconuts.txt (data of Table 11.1)

accession	fruit.weight	fruit.length	husk.thickness	husk.weight
SSAT	984.502	6.875	2.250	245.8
SSAT	1040.00	32.500	2.725	329.0
SSAT	712.00	25.875	2.250	192.8
SSAT	1100.25	28.750	2.475	280.0
POLT	765.25	27.875	3.650	270.0
POLT	713.67	29.500	3.900	333.3
POLT	669.50	28.125	3.425	271.5
POLT	629.50	29.250	3.600	272.5
MVT	1591.67	33.833	3.167	373.3
MVT	1589.25	32.250	2.925	384.3
MVT	2372.50	35.250	4.225	893.5
MVT	1723.25	33.500	3.300	502.0
KKT	1407.50	32.250	2.600	401.8
KKT	1863.75	34.125	3.550	615.0
KKT	1069.50	29.375	2.875	328.3
KKT	1395.50	31.500	3.225	475.5
NLAD	980.75	30.500	3.225	413.3
NLAD	963.50	31.250	3.062	432.5
NLAD	1047.25	31.500	2.925	410.8
NLAD	1056.50	31.375	3.362	472.3

Subject Index

IPGRI and INIBAP
Operate under the name
Bioversity International

Suppoted by the CGIAR

ISBN-978-92-9043-736-9