METHODOLOGY OF BASIC AND APPLIED RESEARCH

Edited by:
A. I. Olayinka, V. O. Taiwo, A. Raji - Oyelade and I. P. Farai

Published by:
The Postgraduate School
University of Ibadan,
Ibadan.

E-mail: postgrad@mail.ui.edu.ng
Website: http://www.postgraduateschool.ui.edu.ng

First Published 2005

ISBN 978 – 37168 – 9 – 1

Printed by: Dabfol Printers, Ibadan.
0803-449-5517

METHODOLOGY OF BASIC AND APPLIED RESEARCH

Edited by

A.I. Olayinka,
V. O. Taiwo,
A. Raji-Oyelade and
I. P. Farai

Contributors

G. Adewale, MEd, PhD. Research Fellow, Institute of Education, University of Ibadan. (e-mail: gbengaadewale@-yahoo.co.uk)

B. O. Agbeja, MSc, PhD. Lecturer in Forest Resources Manage-ment, University of Ibadan. (e-mail: olasinaa@-yahoo.com)

A. J. Ajuwon, MPH, PhD. Senior Lecturer in Health Promotion and Education, College of Medicine, University of Ibadan. (e-mail: ajajuwon@yahoo.com)

T. O. Alonge, MBBS, MD, FRCSEdin, FWACS. Senior Lecturer in Surgery, College of Medicine, University of Ibadan. (e-mail:alonget2003@yahoo.com)

A. O. Arowojolu, MBBS, FMCOG, FWACS, FRCOG. Senior Lecturer in Obstetrics & Gynaecology, College of Medicine, University of Ibadan. (e-mail:ayo_arowojolu@-yahoo.com)

O. C. Aworh, PhD, FNIFST, MIFT, AIDFST. Professor and Head of Food Technology, University of Ibadan. (e-mail:charles.aworh@mail.ui.edu.ng)

J. B. Babalola, PhD. Professor of Educational Management, University of Ibadan. (e-mail:joelbabalola2000@yahoo.co.-uk)

E. A. Bamgboye, PhD, MIS, PSS. Professor and Head of Epidemiology, Medical Statistics & Environmental Health, College of Medicine, University of Ibadan.(e-mail: folabibam@skannet.com)

Contents

7

STATISTICAL ANALYSIS AND INFERENCES

Bamgboye, E. A, Lucas, E. O., Agbeja, B.O., Adewale, G. Ogunleye, B.O. & Fawole, I.

INTRODUCTION

Statistics is a vital tool in any research. Its use starts from the point of gathering data, through data analysis to the point of making the final decision or inferences. The aim of this chapter to take us through these different stages in the use of statistics in research methods and the ultimate objective is to optimize the gains of statistical analysis. These gains are achievable through minimization of errors, correct method of data analysis and reasonable interpretation of results. It is assumed that the reader has had some basic training in the fundamentals of statistics as most terms and concepts are not explained in detail.

DATA GATHERING

The two common means of gathering data for a research purpose are routine collection from a source and data generation through surveys and experiments. Data abstracted from a source are called *secondary data* as the researchers are usually not responsible for the original design and collection of the data. The data generation processes in such cases are usually established for specific purposes not necessarily for the purpose for which the investigator or researcher intends to use the data. But since the information is available and relevant to the study objectives of the researcher, it is cheaper to collect data from a source like this. Historians and social scientists will find this source profitable. Sometimes a desktop review of available information may be sufficient for a needs assessment to plan a

programme. For example, a researcher interested in correlating weather conditions with some events will do well to collect weather parameters from the meteorological services. In medicine, such routine sources include the census, vital registration systems, institutions such as health facilities, schools or armed forces, and disease notification systems such as the cancer registries. Unfortunately, the data from these routine sources suffer from the problems of timeliness, relevance, incompleteness, and inaccuracies.

The second source of data is surveys or experiments. These are planned studies to collect information relevant to specific objectives, including testing of specific hypothesis. The students from the Humanities and the Social Sciences, Public Health, Law and Education are most likely to be involved in carrying out surveys while students of Agriculture, Physical and Biological Sciences, Basic Medical sciences, Veterinary Medicine, Pharmacy and Clinical Medicine are most likely to be involved in conducting experiments.

Another important consideration is the type of research methods which could be either qualitative or quantitative. But since the majority of students will be involved with quantitative research, our major focus will be on guidelines for the analysis of data from this type of research. Regardless of the type of data gathering mechanism, the source yields a primary data. In subsequent sections, the analysis of data from qualitative research shall be discussed.

DATA MANAGEMENT

Data management starts from the onset of data collection. There is no amount of statistical treatment that can completely rectify a badly collected data to make it useful. If for example, certain key variables relevant to the study objectives are not included in the data collected, there is no amount of data imputation or any statistical manipulation that can make the results of analysis adequate and reliable. Also, an error introduced at the data collection stage will propagate through the entire data management process. This may lead to a false inference. It is therefore important to:

- check on the instruments for data collection for completeness of relevant variables to the study objectives;
- ensure adequate training of interviewers or data collectors;
- if necessary, appoint supervisors to monitor data collection;
- monitor completeness and accuracy of data collected and correct any error at the point of data collection.

The best way to do this is to check data collected on a daily basis for immediate detection of errors and corrections. Even when data is collected from routine sources, the researcher needs to check the data for missing values or copying errors. One should ensure that the available relevant data is correctly abstracted from the existing records.

Another step is for the student to decide on the mode of data analysis. This can be either manual calculations or with the aid of the computer. Students are advised to take the advantage of computer technology and use the numerous computer analysis packages which are now available. The computer approach does not only ensure the accuracy of the statistics derived from the data, its results are faster and more comprehensive than the best manual computation method.

DATA CODING/COMPILATION

The first step in the preparation of data for statistical analysis is coding of variables and compilation of records. Coding is the translation of responses to items of information on the questionnaires or data collection sheets to specific categories for the purpose of analysis. This involves the assignment of numbers to the various levels of the variables under consideration. For pre-coded questionnaires, the load of work here is light but certainly important and tedious for open-ended questions. For example, the variable *gender* may be coded with the assignment of 1 to *males* and 0 to *females* while abilities of students to perform certain educational tasks rated as *Low, Medium* and *High* Abilities may be assigned numeric codes 1, 2, and 3, respectively. Coding can also mean the analysis of

factual response data particularly in qualitative research and the subsequent assignment of individuals to classes or categories, or the assigning of categories to individuals.

DATA CLEANING/EDITING

As earlier mentioned, data cleaning should start right from the stage of data collection, when students are advised to go through each record of data collected on a daily basis. The measurement and interview errors, inconsistencies, and inaccuracy of information can be detected at this stage. The checks are also for missing values, copying errors such as writing '80' instead of '18'. Most data sets contain one type of error or the other. After data entry into the computer or into any similar device in readiness for analysis, the data should be appropriately cleaned of a number of errors. Wrong conclusions can be drawn from any data, just because errors have not been checked for and removed from the primary data. This will defeat the aim of the research. Any amount of time spent to ensure that the right image of the data is analysed is therefore justified. Often, once a clean data set is achieved, the analysis itself is quite straightforward and fast, compensating for the time spent in data cleaning. Data Cleaning is a two-step process including detection and then correction of errors in the data set.

Common sources of error during data entry include:

- missing data which have been coded as "999"
- 'not applicable' or 'blank' which have been coded as "0"
- typing errors during data entry
- column shift in which data for one column are entered under the adjacent column
- fabricated or contrived data deliberately introduced
- coding errors

DETECTION OF ERRORS IN DATA SETS

Descriptive Statistics

The calculations in some descriptive statistics can lead to the detection of some types of errors in a data set. Such statistics include frequency counts, missing cases, minimum and

maximum values as well as means, medians, mode, range and standard deviations. With their calculations, one will be able to detect some extreme or impossible values. If for example we assign *male = 1and female = 2* in a data set involving gender distribution, the presence of 3 and 4 in the data set will imply error in coding or in data entry. Also, if the standard deviation is higher than the mean value, then some impossible extreme values or outliers must have been included in the data set. When these are fished out and corrected, the standard deviation will become smaller. The following are a few of the ways by which errors can be fished out.

Graphs

Scatter plots. One other way to inspect the data for errors is to plot one set of variables against another on a simple x – y graph. This may reveal the relationship between the two sets of variables under consideration. If one takes a close look at the general trend, one may see some points that do not conform to the general trend. Such points may need rechecking. Plotting the scatter plots will show the outliers, that is, points which do not follow the general pattern of association between the two sets of variables.

Histograms A histogram is the graphical display of frequency distribution in form of rectangular bars whose heights represent the frequencies of the data intervals represented by their widths. A histogram can also reveal extreme values in a data set by showing values which are not within the range of the distribution, especially when normal distribution is assumed. If the data set is small, i.e. less than 100 cases and 10 variables, cleaning can be done once and for all. However, the process is done in several stages for a large data set. It begins with the examination and correction of the key and important independent variables (e.g. treatment, gender and age). Other variables may be cleaned as each of them become relevant in the analysis. The most important thing is never to introduce new variables into an analysis without first checking for errors and making corrections.

DATA ANALYSIS

Statistical quantities employed in data analyses are used for either descriptive purposes or for drawing of inferences or both. Although there are some special statistical methods for some specific data situations and statistical questions, there are some common standard statistical procedures which cut across all disciplines. These procedures are useful for carrying out univariate analysis, bivariate analysis or multivariate analysis. The following are the initial steps basic to all statistical analyses of data:

- Identify items of information relevant to the study objectives.
- Identify independent or explanatory variables.
- Identify response or dependent variables.
- Identify variables to be derived from other variables.

The computer approach is most favoured for data analysis; but in doing this, the researcher should check on the following concerns:

- Is data qualitative or quantitative?
- Is data already in computer format?
- Any need for post coding of data?
- Which statistical package will be used for database creation?
- Any need to consult a statistician or expert on the subject?

As a general rule, the type of data analysis to carry out depends on the following:

- Study objectives
- Type of data
- Design of study - any effort to reduce variability in sets of data or minimize errors
- Nature of samples - independent or dependent
- Sample size

General format of statistical analysis of data

Statistics is used specifically to

- reduce a large quantity of data to a manageable size,
- present data in understandable form,
- aid in the study of populations,
- aid in decision making, and
- aid in making reliable inferences from observed data.

The following are necessary common activities as first steps in the statistical analysis of quantitative data:

- Make frequency distribution tables for all variables.
- Compute appropriate descriptive statistics.
- Make necessary cross-tabulations dictated by study objectives.
- If comparing or looking for associations among variables, plot data in a scatter diagram and apply the appropriate test statistic to determine strength of association.
- Choose appropriate statistical tests to use, which could be either parametric or non-parametric. The choice of which to use is indicated by how much the data available satisfies the assumptions underlying the developments of the statistical tests.

As indicated in earlier sections the first activities in data analysis are as follows:

- Prepare a coding format (for example if the statistical package-Epi-Info is to be used, there is the need to design the data entry questionnaire).
- Post code all open questions.
- Ensure all variables are coded appropriately.
- Prepare dummy tables.
- Start data entry.

Nature of Statistical Analysis

The nature of statistical analysis depends on the research questions which dictate the study objectives in the first instance and which inform the type of data collected. The usual objectives of most studies arise from the need

- to estimate certain population parameters,

- to compare attributes of data in many groups,
- to determine the best treatments or interventions to produce certain results,
- to find relationships between variables to explain observed effects, and/or to predict future events from observed data.

The researcher should be able to identify the extent of statistics required for data analysis. He should also know if his study will only need descriptive statistics or will require statistics to test a certain hypothesis and therefore draw appropriate inferences from the data. It is not every statistical work that requires a test of hypothesis or that must contain P-values or some of the statistical tests of significance as erroneously believed by some researchers. Such researchers, who are often graduate students driven by the desire to impress the examiners, usually carry out intervention studies even at the expense of available resources such as time, finance and manpower. This notion is not correct. Researchers should know that statistics is only a means and not the end in itself. The research objective should dictate the extent of statistical analysis required.

Common Statistical Procedures

Descriptive Statistics

These include frequency counts, minimum and maximum values as well as measures of central tendency such as the means, medians, mode, range and measures of dispersion such as standard deviations. Other statistics are ratio, proportion, percentages and rates.

Use of Graphs or diagrams

For evocative displays of data to draw attention to the salient features and patterns in the values, graphs and diagrams are used to summarize and present data. These graphs or diagrams include scatter diagrams, histograms, pie charts, bar charts, pictograms and line graphs. Some of these displays such as scatter plots and histograms may as previously discussed, reveal certain features that may suggest the need for cleaning of data before further statistical analyses. The choice of appropriate

descriptive statistics, graphs or diagrams depends on the types of data collected and the objective of the study.

TYPES OF DATA

Let us quickly remind ourselves of the two types of data available. These are qualitative and quantitative data.

Qualitative Data

Qualitative data arise when the observation fall into separate distinct categories with no notion of numerical magnitude. Such data are measured on the nominal or ordinal scales. Recall that nominal scales are mainly classificatory and there is no natural order between the categories which are also mutually exclusive, as no individual can belong to more than one category. Such data are inherently *discrete*, in that there is a finite number of possible categories into which each *observation* may fall. Examples include:

- Colour of eyes: blue, green, brown and white;
- Examination result: pass or fail;
- Type of medical diagnosis: Hypertension, diabetes mellitus, cancer of the breast, HIV infections, asthma, tuberculosis and irritable bowel syndrome;
- Gender: Male, Female.

In the ordinal scale of measurements, an ordering exists as the mutually exclusive categories are graded. It is sometimes referred to as ranking scale. Examples include:

- Level of Educational Qualification: Teachers Grade II certificate, NCE, B.Ed. or B.A., Masters, and PhD degrees;
- Socio-economic status: low, middle or high; Level of pain: mild, moderate, severe.

Data collected on these scales are referred to as categorical data.

Quantitative Data

This type of data has the notion of numerical magnitude. In other words, the values are expressed in numbers and in some

cases the units of measurements are well known. They have all the properties of nominal and ordinal scales but are measured on at least the interval scale. In the interval scale of measurement, the zero level is always arbitrary but the differences between successive points are equal. That is, the difference between 80% and 81% is the same as the difference between 81% and 82%. Examples are: Students' scores in a school examination and "Temperatures of patients measured on either Celsius or Fahrenheit units". The scale has both numerical magnitude, direction (interval) and an absolute or true zero. For example, height, weight, and age all have absolute zeros regardless of their units of measurements. For example, zero centimetre equals zero feet if height is measured on either unit.

The data on these scales of measurements are said to be discrete if the measurements are integers assuming only whole numbers or counts. Examples are the number of students in a class, parity, and apgar score. They are continuous if the measurements can take on any value, usually within some range in a continuum. For example, students' score in a Biology test is a discrete variable while weight and skin fold thickness are continuous variables.

TYPES OF ANALYSIS

Univariate Analysis

This is when each variable is examined at a time in the analysis to examine the distribution of values of each set of variables. This task is also called data exploration. The statistics appropriate in this situation is to provide precise descriptive analysis and to make frequency distribution tables for all variables.

Frequency distributions

These are used basically to describe a given set of data according to the number of times a value or a data interval occurs in the data set. They can also be used to test whether two or more distributors are sufficiently similar to warrant merging them. For instance, if one is studying the performance of boys and girls in English Language, distribution of the scores of boys

and girls may be similar. If this is so, the data for boys and girls can be merged for analysis.

Graphs

Graphs are two-dimensional representations of relations between pairs of variables. If a relation or interaction exists in a set of data, a graph will show it as well as define its nature. The types of relationship between two variables include linear, quadratic exponential, etc.

Descriptive Statistics

Descriptive Statistics include measures of central tendency as well as measures of variability or dispersion. The three common quantities for measuring average or central tendencies or location are the *mean, median* and *mode*. They tell what the data 'look like' on the average. The *mean* is the most important measure of central tendency, and the most widely used in research. However, the *median* (the value in the middle when the data is arranged in an increasing or decreasing order) and the *mode* (the mostly occurring value in a set of data) can sometimes be useful in some data situations. In particular, the median is more appropriate than the mean for graded responses or quantitative data having skewed distributions. For example, the *median* is used as a cut-off point or mark to rank subjects on a set of scores in categories or classes e.g. high and low abilities.

On the other hand, the measures of variability include the range, variance and standard deviation. They are very important because adequate interpretation of data is virtually impossible without a good knowledge of the variability in the data, as measured by the standard derivation. In fact, means as summary indices are not usually reported without the standard deviations. And when one reports the median as a measure of average, one should also report the range.

Bivariate Analysis

This is indicated when one is interested in examining the relationship between two variables simultaneously. One set of variables is termed dependent variable and the other, independent variables. Such analysis can be used to compare

two groups of data based on certain indicator variables and in each case test certain hypothesis. In particular, the interest may be to look for the effect of certain treatments on some physiological and biochemical parameters. Among other possibilities, the bivariate system of analysis can examine the effect of certain teaching methods on the performance of students at an examination or find out the effect of two different fertilizers on the yield of certain agricultural crops. The rationale behind this kind of analysis is to allow us take decisions. Thus, it is important to note the procedures for testing hypothesis.

Tests of Significance

A test of significance is a statistical test that attempts to determine whether or not an observed difference indicates that the given characteristics of two or more groups are the same or different; or whether a relationship exists between two or more variables. The procedure for investigating the truth of any hypothesis is called hypothesis testing and there are six steps by which the tests are accomplished.

Step 1. State the Null Hypothesis

The hypothesis is stated in a form that it would be nullified if available data do not support it. If it were to compare the difference in means between two groups, the null hypothesis will be stated as: There is no difference between the means of the two groups. If the problem were to test for associations, then the null hypothesis will assume that there is no association between the categorical variables. In a study to determine the effect of a drug suxamethonium on serum potassium levels, the null hypothesis is: the drug suxamethonium has no effect on the serum potassium levels. It is clear that we are always testing the null hypothesis and we only calculate the probability of observing a value as extreme or as more extreme than observed for our test statistics if the null hypothesis were true.

Step 2: State the Alternative Hypothesis

This is usually the research question and it is the hypothesis to fall back on whenever the null hypothesis is rejected. This can be stated either in one direction to yield a one-tail test if one is

sure of the direction, or in a two-way direction to yield a two-tail test. For example, the alternative hypothesis to the one stated above is, can the drug suxamethonium have an effect on the serum potassium levels?

Step 3: Set the criterion for rejection

How different must the means be before we can say the mean of group A is not the same as that of group B? Or how big must the effect be before we can say that it is significant or not significant? The probability of wrongful rejection of the null hypothesis should be very small. By convention, a probability of 5% is chosen. This implies that there is only a 5% chance or less that the mean of group A is similar to the mean of group B. This 5% describes the area of the unit normal curve with 2.5% on the right and 2.5% on the left sides of the curve. The level is called level of significance or the alpha (α) level. Although 5% is very common, we can set α level to more stringent values such as 0.001, 0.01 or less stringent values such as 0.10 or more.

These values correspond to the probability of observing such an extreme value corresponding to the standard normal deviate by chance. Another interpretation of the significance level α, based on decision theory, is that α corresponds to the value for which one chooses to reject or accept the null hypothesis H_o.

Step 4: Choose appropriate test statistics

Recall that you are at liberty to use a parametric or non-parametric test statistic according to whether you are assuming an underlying distribution for your data or using a distribution-free method. The relevant parametric test statistics for comparing two mean values is the student t-test, if the two groups are independent; otherwise it is the paired t-test if groups are dependent. If your data do not satisfy the assumptions underlying the use of this test statistics, then a non-parametric equivalent is the Mann-Whitney-U test if the groups are independent or the Wilcoxon-rank sum test if dependent.

Step 5: Evaluate the test statistics

Although this can be done manually for a small data set, the use of the computer is recommended for reasons stated earlier.

Indeed there are many statistical packages that are available for calculating virtually all the test statistics depending on the problems.

Step 6: Conclusions and drawing of inferences

This last step in the test of hypothesis entails drawing appropriate conclusions from the observed data and taking decisions. Recall that the p-value or α- level of error is the probability of observing a calculated value of the test statistic or an extreme value if the null hypothesis were true. These p-values or α- level of errors can be obtained from appropriate statistical tables to test the statistics used. For example, there is a table for t-distribution from where the appropriate level of error can be obtained if the student t-test has been used. Recall that the student t-test is appropriate for comparing two mean values. The decision rule is based on the α- level error. If the value of t from the table at the chosen α- level error is smaller than the calculated value in step 5 above, we reject the null hypothesis and take the alternative hypothesis. It is very common to choose α- level of 5% even though other levels can be chosen. Then we say if $p<0.05$, we reject the null hypothesis and conclude that our data do not support the null hypothesis.

One common problem is to generalize our result beyond the sample data. There are a lot of considerations before generalizing results beyond the study sample. You may need to check how representative of the target population your data is. How big are your sampling errors? How well have you taken care of confounding factors. Can you strictly attribute observed treatment effects to the treatments only? In other words, a lot of caution is necessary when interpreting findings. As much as possible, make sure to write the report in the context of the problem.

SOME OTHER STATISTICAL TESTS

Let us now describe the most commonly used statistical tests for measuring the strength of association between two variables or comparison of two groups in detail.

Test of Association

It may be the objective of the study to examine the association between two variables. The strength and weaknesses of this association could be very important as one of the steps to determine causal relationship. As mentioned, the type of data is one of the factors that will determine the type of statistical analysis and test statistics.

Qualitative Variables and the Contingency Table

The values of qualitative variables are attributes which merely define the categories of interest. These attributes are mutually exclusive and each individual subject can only belong to one category of the variables. In examining the association between any two categorical variables, each subject is allocated to corresponding categories. The resultant cross-tabulation yields contingency tables whose size is dictated by the number of categories of each variable. Note that the entries in each cell of the tables are numbers and the contingency table consists of a number of rows and columns. The smallest contingency table is a 2×2 contingency table in which each variable has only two values and the contingency table yields two rows and two columns.

Example

The interest of the researcher may be to find out the association between survival rate and the type of treatment given for a particular health condition. The patients are divided into two groups; group A was given streptomycin as well as bed rest and the other group B placed on bed rest only. The results showed that 4 out of 55 patients in group A died before 6 months period while 14 out of 52 in group B died during same period.

Question: Is there any significant association between treatment received and survival?

Organize data in a 2x2 table

	Outcome of Treatment		
Treatment	**Death**	**Survival**	**Total**
Group A	4 (9.25)	51 (45.75)	55
Group B	14 (8.75)	38 (43.25)	52
Total	18	89	107

Two contingency tables have been combined above. One is from the observed data given while the other is calculated based on the null hypothesis that there is no difference between the two methods of treatment. The data for the calculated is given in parentheses in each cell. The statistical test to be carried out involves comparing the calculated value of χ^2 (Chi-square) based on the difference between the two contingency tables with the value of χ^2 read from the table at $\alpha = 5\%$ and $(r-1)(c-1) = 1$ degree of freedom. The calculated χ^2 is based on the following definition:

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} = \frac{(4-9.25)^2}{9.25} + \frac{(14-8.75)^2}{8.75} + \frac{(51-46.75)^2}{46.75} + \frac{(38-43.5)^2}{43.5} = 7.37$$

where O_i is the observed frequency in the i^{th} cell of the contingency table and E_i is the expected frequency given that the Null Hypothesis is true. The expected value under the null hypothesis that "there is no association between type of treatment and outcome of treatment" is given in brackets in the table. The value of χ^2 from the table at 1 degree of freedom at 5% level is 3.841, which is far less than the calculated value of 7.37. Therefore we conclude that there is a statistically significant association between type of treatment and outcome of treatment ($P<0.05$). A look at the data shows that the survival rate is better in the group given streptomycin in addition to bed rest.

Yate's correction for continuity

When we have a 2 × 2 contingency table as we have above, we should make some corrections to allow for the discrepancy in approximating a discrete sampling distribution with a continuous distribution. This correction is accomplished by subtracting ½ from the absolute difference between the observed and expected frequencies. That is:

$$\chi^2 = \frac{\sum(|O_i - E_i| - \frac{1}{2})^2}{E_i}$$

The quantity $\frac{1}{2}$ is Yate's correction for continuity. The researcher can try and find the corrected value of χ^2 in the above example.

Paired Samples

One of the criteria to consider in the choice of a test statistic is the design of the study. We check whether the samples are dependent or independent. Recall that to ensure comparability of groups, one way is to match groups on important variables and we only attribute any differences to the interventions. Sometimes, the same subjects are exposed twice and the differences are measured. An example will make this clearer and we now consider qualitative variables.

Example

A retrospective study was done to find out if certain strains of influenza could be teratogenic when infection occurs during the first trimester of pregnancy. Each of 229 women who had defective babies following an influenza epidemic was matched with a control among those who had normal babies following the epidemic. The data from this design revealed that there are four different types of women. The result is as follows:

Type of Pairs	Women with defective babies	Women with no defective babies	Number of pairs of women
1	+ve influenza	+ve influenza	80
2	+ve influenza	-ve influenza	40
3	-ve influenza	+ve influenza	63
4	-ve influenza	-ve influenza	46

This data can also be presented in a 2 × 2 contingency table as shown below

	Influenza +	influenza -	
Birth +	80®	63(s)	143
Defect -	40(t)	46(u)	86
	120	109	229

Using Mc –Nemers' Test statistics; $X^2 = \frac{(s-t)^2}{s+t} = \frac{(63-40)}{63-40} = 5.14$

P<0.05

Conclusion

The babies without birth defects appear more likely to be free from influenza. You may need to consult appropriate Statistics textbooks for details of the statistical methods chosen.

Further use of the Chi-Square Test

Goodness of Fit Test

Often, we may need to compare an observed frequency distribution with the expected distribution from a theoretical model. We may also want to test an observed frequency distribution against a known population frequency. The statistical procedure of test of goodness of fit using χ^2 test becomes very useful in doing this.

Example

Gastric Cancer is postulated to occur more often in persons with Type A blood. Suppose 200 patients with gastric cancer have the following percent of blood type: O(38%) A(52%), B(1%) AB (3%) and suppose it is established in the general population

that the percents of AOB groups are: O(45%), A(40%), B(12%) and AB(3%).

Question: Is there a statistically significant difference between the blood type frequencies in the sample and in those existing in the general population?

Solution:

	Blood Groups				
	O	A	B	AB	Total
Observed Frequency	76	104	16	4	200
Expected Frequency	90	80	24	6	200

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} = \frac{(76-90)^2}{90} + \frac{(104-80)^2}{80} + \frac{(16-24)^2}{24} + \frac{(2-6)^2}{6} = 12.71$$

The value of χ^2 from the table at $v = n - 2 = 2$ degrees of freedom at 5% level is 5.99, which is far less than the calculated value of 12.71. The conclusion here is that there appears to be a statistically significant difference between the distribution of blood types among patients with gastric cancer and the distribution in the general population.

Fisher's Exact Test

This is applicable in a data situation when there are extremely low frequencies in a 2 × 2 contingency table and one is in doubt as to the adequacy of continuity correction. In this case, it will be necessary to calculate the probability of observing the figures in the cells conditional on the marginal totals.

Example

Efficacy of Anti spasmodic drug in controlling seizures

	Seizure Status		
Drug	+ ve	ve	total
Placebo	2 (a)	6(b)	8 (a+b)
	7 (c)	1 (d)	8 (c+d)
Total	9 (a+c)	7 (b+d)	16 (n)

$$P = \frac{(a+b)!\,(c+d)!\,(b+d)!\,(a+c)!}{n!\;a!\;b!\;c!\;d!}$$

$$P = \frac{8!\,8!\,9!\,7!}{16!\,2!\,1!\,6!\,7!} = 0.020$$

REGRESSION ANALYSIS

If the researcher's interest is to determine the relationship between variables in terms of magnitude and direction, a regression analysis or correlation analysis is carried in the case that the data are quantitative. When the objective is to explain one variable from the knowledge of the other for purposes of prediction, the regression model is always fitted to the data. But if it is to determine the strength of relationship between the variables, the correlation coefficients are calculated. Again, the choice of test statistics to determine the strength of association or relationship between the variables depends on the kind of data and this will suggest the use of either parametric or non-parametric statistics. Thus, the product-moment coefficient of correlation (r) or the Spearman rank-order correlation coefficient (r_s) may be calculated as a measure of relationship. The value of the coefficient lies between -1 and +1 and the student t-test can be used to assess its statistical significance. Please note that when only two variables are involved, the product-moment correlation coefficient is mostly used provided the data set satisfies the assumptions underlying the use of a parametric test.

Some Uses of Regression Analysis

1. To learn if Y, the dependent variable, does depend on X, the independent variable;
2. To predict Y from X;
3. To determine the shape of the regression curve;
4. To determine error in Y in an experiment after adjustments have been made for the effect of a related variable X;
5. To test a theory about cause and effect.

Steps in Regression Analysis

Regression analysis is performed on populations with bivariate distributions i.e. distributions in which the variables are associated in pairs. That is, for every measurement of a variable X, there is a corresponding value of a second variable Y. The variable from whose values predictions are to be made is denoted by X and variable whose values are to be estimated or predicted is denoted by Y.

1. The first step in regression analysis is to represent the n pairs of values of X and Y as n points on a graph to obtain what is commonly referred to as a scatter diagram. The independent variable X is plotted along the horizontal axis while the dependent variable, Y is plotted along the vertical axis.

2. A straight line, the line of "best fit" is drawn connecting as many points as possible on the scatter diagram. This line is called the sample regression of Y on X. Its position is fixed by two results:
   (i) It passes through the point O (X, Y), the point determined by
   (ii) the mean of each sample.
   (iii) Its slope is equal to b in the unit of Y per unit of X, where b is the coefficient of X in the regression equation..

Now, the equation of any non-vertical line can be written in the form: $Y = a + bX$ where a is the intercept on Y-axis and b the slope of the line. A particular line, therefore, is completely determined if the values of the constants a and b in its equation are known. Therefore, in order to find the line of best fit to a scatter diagram of n points, we must determine a and b in such a way that the n points lie as close to the line as possible. After the equation of the line of best fit has been determined, it will yield for each X-value a certain Y-value, which will be an estimate of the actual Y-value. The equation of the line of best fit can therefore be written in the form

$Ye = a + bX$ where Ye is the estimated value obtained from the line and Y is the actual value obtained by measurement.

Example

The age X in weeks and mean height Y in centimetres, of cowpea plants are given in the following table. Find the equation of the line of regression of Y on X. Calculate the standard error of estimate for the regression.

X	Y	XY	X^2	Y^2	Y_e	$Y-Y_e$	$(Y-Y_e)^2$
1	5.5	5.5	1	30.25	-0.52	4.98	24.800
2	6.8	13.6	4	46.24	13.19	-6.39	40.800
3	25.0	75.0	9	625.0	26.9	-1.9	3.61
4	40.0	160.0	16	1600	40.61	-0.61	0.372
5	55.0	275.0	25	3025	54.32	0.68	0.462
6	72.0	432.0	36	5184	68.03	3.97	15.761
7	80.0	560.0	49	6400	81.74	-1.74	3.028
28	284.3	1521.1	140	16910.49			88.865

The mean $\bar{X} = 4$ while the mean $\bar{Y} = 40.614$. The regression equation $Y = a + bX$

$$b = \frac{\sum XY - \bar{X}\bar{Y}}{\sum X^2 - n\bar{X}} = \frac{1521.1 - 7(4 \times 40.614)}{140 - 7 \times 16} = 13.71$$

$$a = \bar{Y} - b\bar{X} = 40.60 - (13.37 \times 4) = -14.23$$

That is, the equation of the regression line is therefore

$Y = -14.23 + 13.71X$

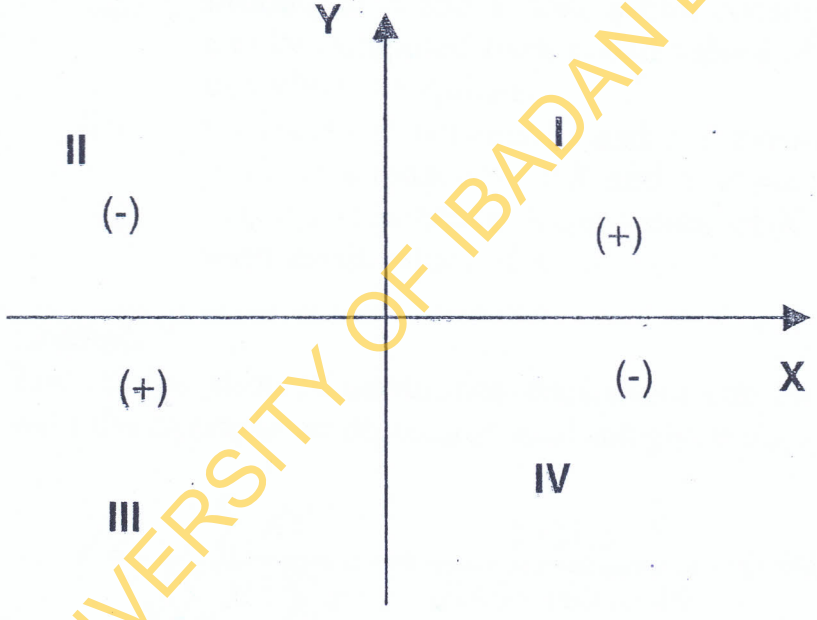
Standard error S_e of estimate for the regression n calculated is given by: $S^2{}_e = \frac{\sum (Y - Y_e)^2}{n}$

where Y_e is the predicted value of Y from the regression equation. That is,

$$S_e = \sqrt{\frac{88.865}{7}} = 3.56$$

CORRELATION ANALYSIS

Correlation coefficient is another measure of the mutual relationship between two variables. Correlation coefficient denoted by r is a quantitative expression of the strength of the linear relationship between two variables. The relationship between the variables X and Y can be determined from a sample of n pairs and the first step is the construction of the scatter diagram for the data to illustrate the existence and the nature of a relationship. The figure below shows the signs of the cross products, XY in the four quadrants of x-y coordinate system. If in each quadrant, the number of points in the scatter diagram is approximately the same, no linear relationship is indicated, and the sum of the products ΣXY will be numerically small.



If, however, the points follow a linear trend and occur mostly in quadrants I and III, the positive products will predominate, and ΣXY will be positive; if most of the points fall in quadrants II and IV, ΣXY will be negative. It is evident therefore, that the sum of the product (ΣXY) and the mean product (ΣXY/n) are indicators of a linear relationship between the values of X and Y, and that the larger ΣXY is in absolute value, the closer the points lie to a straight line and the stronger is the evidence of a linear relationship.

To find the sample correlation coefficient, denoted by r, we compute Σx^2, Σy^2 and ΣXY as was done for regression analysis. Then

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

The sample correlation coefficient r is a measure of the degree of the linear relationship between two variables.

Two Properties of r

(i) r is a number without units or dimensions, because it is a ratio of two quantities that are of the same product of X and Y. One useful consequence is that r can be computed from coded values of X and Y. No decoding is required.

(ii) r always lies between -1 and +1. Positive values of r indicate a tendency of X and Y to increase together. When r is negative, large values of X are associated with small values of Y.

Example

The computation of correlation coefficient can be demonstrated with the example on regression analysis given above.

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \frac{1521.1}{\sqrt{140 \times 16910.49}} = 0.99$$

This implies very strong correlation. The significance of the correlation coefficient can be examined by postulating a Null Hypothesis that there is no correlation between X and Y. That is, the population coefficient is 0. The statistic to test is given by:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.99}{\sqrt{\frac{1-(0.99)^2}{7-2}}} = 15.69$$

For n = 7 or degree of freedom v = 5, the value of t from the student's t-distribution table at 5% level, $t_{0.05}$ = 2.015. That is, the calculated t is greater than t from the table. We therefore reject the Null Hypothesis and accept the alternative. That is, the correlation coefficient of 0.99 is significant meaning that age of the plant (X) is linearly related to the height of the plant (Y).

COMPARISON OF TWO GROUPS: QUANTITATIVE DATA

Analysis of Differences

A research study may seek to analyse differences between means. For instance, if the objective is to ascertain the effect of some forms of instruction on pupils' performance, the difference between the variances of groups taught by differing methods can be tested easily. The t-test is relevant if only two groups are involved in the experiment while the Analysis of Variance (ANOVA) becomes relevant with more than 2 groups. Since the t-test and ANOVA are both inferential statistics, certain assumptions underline their use. One of these assumptions is that the observations (scores) are independent and the value of any one observation should not be related to the value of another observation. Another assumption is that the scores in the population are normally distributed.

Also, the Analysis of Covariance (ANCOVA) invented by Ronald Fischer is appropriate when the subjects in two or more groups are found to differ significantly on a pre-test or other initial variable. In this case, the effects of the pre-test and/or other relevant variables are adjusted and the resulting adjusted means of the post-test scores are computed.

The Student T-test

The student t-test has an underlying normal distribution and, as earlier mentioned, it is effective in situations where the sample size is less than 50 and the population standard deviation is

estimated from the sample data. The major difference between the normal distribution and the student t-distribution is that the latter has more areas at the tails of the curve which is dependent on the degrees of freedom. Therefore a new Table of areas was constructed for the student t-distribution and simply called the t-table. The standard t-table is based on *n-2* degrees of freedom and when the sample size is more than 50 the t-table and Z-table have the same areas under different segments of the curve. In many clinical researches, the sample size is usually small and the values of the population standard deviation unknown. This explains why the student t-test is one of the most popular statistical tests with medical doctors when comparing two mean values.

$$s^2 = \frac{\sum(x_1 - \bar{n})^2 + \sum(x_2 - \bar{x}_2)}{n_1 + n_2 - 2}$$

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

We start the t- test by setting a Null hypothesis that there is no difference in the means of the two samples being compared. That is, they are from the same population, which will imply that $\mu_1=\mu_2$. The test statistic, simply written as t-test is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}$$

The number of degrees of freedom is given by $(n_1+n_2)-2$ while $\bar{x}_1$ is the mean of the measurement in the first sample with size n_1, $\bar{x}_2$ the mean of the measurement in the second sample with size n_2 and $s_1$and s_2 are the standard errors of the samples. When the t-value is evaluated, the probability of

obtaining a value as extreme as that obtained under the null hypothesis is derived from the t-table at the appropriate degree of freedom.

Pooled Variance

Suppose we have two samples x_1 and x_2 of sizes n_1 and n_2 which have been drawn from two populations whose variance s^2 is unknown. The pooled variance of the two samples can be calculated.

Example

The mean time taken in days to regain birth weight by 20 children fed on the standard premium formula is 13.1 days with a standard deviation of 5.3 days. Another comparable group of 20 children fed on prematalac milk had a mean of 9.9 days and a standard deviation of 3.7 days to regain their birth weight. Is there a significance difference in the number of days to regain birth weight as a result of the different types of milk? Comment on your findings.

Solution:

Step 1: Set up a Null Hypothesis which states that there is no difference in the number of days taken to regain birth weight between the children fed on the standard premium and prematalac milk. That is, the two samples are from the same population.

$$H_o : \mu_1 - \mu_2 = 0$$

Step 2: Set up an Alternative Hypothesis that there is a difference between the means and hence the samples belong to two different populations.

$$H_A : \mu_1 - \mu_2 \neq 0 \quad \text{or}$$
$$H_A : \text{Either } \mu_1 > \mu_2 \text{ or } \mu_1 < \mu_2$$

That is, no direction of the difference has been given and so we call this scenario a two-tailed test. But the second expression for H_A when a direction is given, we call a one-tailed test.

Step 3: The level error (α) for statistical significance is 0.05.
Step 4: The test statistics chosen is the student t-test for independent samples which according to the definition when Null Hypothesis is assumed is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}} = \frac{13.1 - 9.9}{\sqrt{\frac{5.3^2}{20} + \frac{3.7^2}{20}}} = 2.214$$

Step 5: Check the t-table for t_T at $\alpha = 0.05$ and $v = 20+20-2 = 38$ degrees of freedom. The value of t_T from table is 1.64.
Step 6: We now conclude that our data is not consistent with the null hypothesis and it appears that the difference in the average time to recover birth weight is statistically significant. In other words, it appears children on the new milk prematalac recovered birth weight more quickly. If we had used the pooled estimate variance ($S^{2)}$ in calculating the standard error, the conclusions could have been the same. Thus pooled estimate of variance:

$$s^2 = \frac{19 \times 5.3^2 + 19 \times 3.7^2}{20 + 20 - 2} = 20.89$$

In that case, our calculated $t = \dfrac{131 - 9.9}{\sqrt{20.89\left(\frac{1}{20} + \frac{1}{20}\right)}} = 2.214$

with 38 d.f as before.

COMPARISON BETWEEN TWO MEAN VALUES: DEPENDENT GROUPS

As mentioned earlier, the design of the study may be such that it has some pairing properties where for each value of a variable in one group there is a twin relation in the other group. Such designs results when:

i. matching has taken place on all possible confounding factors that could cause systematic differences or

ii. observations are taken from the same experimental units at two different times say before and after an intervention (treatment). Units serve as self-control.

The student t-test used in this situation is called the **Paired t-test**.

Student paired t-test

The focus of analysis is on the difference between the **individual pairs** of observations rather than the difference between the means of each group. Therefore the mean of the differences between individual pairs is tested rather than the differences between the means. Thus paired t-test is given by the formula:

$$t = \frac{\bar{d} - 0}{\sigma}$$

where $\bar{d}$ is the mean of the differences between individual pairs and σ is the standard error of the differences between each paired observations. Recall the formula for the standard error of a mean as standard deviation divided by the square root of the sample size which is the number of paired observations in this case.

Paired Sample

Ten patients participated in a clinical trial of a new drug to control hypertension. Their systolic blood pressures before and after the use of the drug are as follows:

Patients	1	2	3	4	5	6	7	8	9	10
Before	160	140	180	160	225	150	150	140	170	165
After	140	150	170	130	180	120	150	120	130	140
d_i	20	-10	10	30	45	30	0	20	40	25

Is there any significant effect of the drug?

Solution: You should follow the six steps as in the previous example.

Step 1: Ho: $\bar{d} = 0$

Step 2: H_A: $\bar{d} \neq 0$
Step 3: $\alpha = 0.05$
Step 4: choose student paired-t test.

$t = (\bar{d} - 0) / \sigma$ where $\bar{d} = \Sigma d_i / n = 210 / 10 = 21.0$

$$\text{and } \sigma = \frac{std.dev}{\sqrt{n}} = \frac{17.13}{\sqrt{10}} = 5.416$$

$$\text{That is, } t = \frac{21.0}{5.416} = 3.877 \text{ on d.f.} = n - 1 = 9$$

Step 5: Checking the table for the significance or P-value, we find $P<0.05$.
Step 6: Conclusion: The drug effect is statistically significant and it appears it has a lowering effect on systolic blood pressure.

QUALITATIVE DATA

Parametric test of significance

These are statistical tests whose models do not require strict conditions or assumptions about the form of distribution for the population parameters under test. Hence they are sometimes called ***distribution-free methods***. It is specifically useful to test hypothesis on such data with any of the following characteristics:

i. Few observations
ii. Skewed distributions that cannot be transformed to normal distribution
iii. Graded responses usually on the nominal and ordinal scales of measurements
iv. Quick and easy analysis
v. Results requiring exact probabilities

Test of Means

To compare the average values computed from two independent samples, the non-parametric test to use is Mann-Whitney U Rank-Sum Test or the equivalent, Wilcoxon two sample rank tests. For paired samples, the Sign Test or the Wilcoxon signed Rank tests are used. The Analytic Approach is to use the relative positions of observations rather than their actual values in the ranking. Where the data of interest meet the assumptions

underlying the parametric tests, the non-parametric tests are weaker! These tests will not be described here but can be found in statistical textbooks.

Further Qualitative Data Analysis

Standard statistical techniques exist for the analysis of qualitative data from focus group discussions and in-depth interviews. Recall that such data are usually collected as:

1. Text
2. Audio-taped Conversation/Responses
3. Observed Behaviour
4. Video-taped Behaviour

The method of analysis include

- Text coding;
- Intuition;
- Graphical displays;
- Grounded theory methods;
- Deconstruction;
- Levi-Straussian analysis and;
- Dream analysis.

MULTIVARIATE ANALYSIS

The multiple correlation coefficient (R), Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA) and other multivariate analysis techniques such as multiple regression, canonical correlation, discriminant analysis, factor analysis, path analysis or causal modelling may also be used depending on the research question or hypothesis and data available.

MULTIPLE REGRESSION ANALYSIS

There are data situations when many variables are required to explain the dependent or response variables. In this regards, the regression of Y on a single independent variable is often inadequate. Two or more Xs may be available to give additional information about Y by means of a multiple regression on the X's. Some of the uses of multiple regression analysis are:

i. To construct an equation in the X's that gives the best prediction of the values of Y.
ii. To find the subset that gives the best linear prediction equation when there are many Y's.
iii. To discover which variables are related to Y, and, if possible to rate the variable in order of their importance.

Multiple regression analysis is a complex subject. The calculations become lengthy and cumbersome when there are numerous X-variables, and it is hard to avoid mistakes in computation. However, standard electronic computer programmes are available for such analysis.

Example
Multiple linear regression analysis may be illustrated by an investigation of the source from which maize plants in various soils obtain their phosphorus. Or the effect of certain socio-demographic variables on the values of systolic blood pressures. Results of an experiment to measure the concentrations of inorganic (X_1) and organic (X_2) phosphorus in different types of soil and the phosphorus contents (Y) of corn grown in such soils are shown in the table below these three variables. Let us consider how to determine X_1 and X_2 then determine Y

Soil Sample	X_1 (ppm)	X_2 (ppm)	Y
1	0.4	53	64
2	0.4	23	60
3	3.1	19	71
4	0.6	34	61
5	4.7	24	54
6	1.7	65	77
7	9.4	44	81
8	10.1	31	93
9	11.6	29	93
10	12.6	58	51
11	10.9	37	76
12	23.1	46	96

13	23.1	50	77
14	21.6	44	93
15	23.1	56	95
16	1.9	36	54
17	26.8	58	168
18	29.9	51	99

Solution

In this example, there were 18 soil samples; the multiple regression equation is as given in equation (1)

Prediction equation $Y = a + b_1 X_1 + b_2 X_2$ (1)

where $a = Y - b_1 X_1 - b_2 X_2$ (2)

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)} \quad (3)$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1^2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)} \quad (4)$$

x_1 is obtained by subtracting $\bar{X}_1$ bar$_1$ from each X_1

$x_1 = X_1 - \bar{X}$, $x_2 = X_2 - \bar{X}_2$ and $y = Y - \bar{Y}$

$\bar{X}_1$, $\bar{X}_2$ and $\bar{Y}$ are the means of X_1, X_2 and Y, respectively.

It is advised to construct another table as follows to facilitate the calculation of these quantities.

Soil Samples	X_1	X_2	Y	$y(Y-\overline{Y})$	$x_1(X_1-\overline{X}_1)$	$x_1(X_1-\overline{X}_1)$	x_1x_2
1	0.4	53	64	-17.28	-11.54	10.89	-125.67
2	0.4	23	60	-21.28	-11.54	-19.11	220.53
3	3.1	19	71	-10.28	-8.84	-23.11	204.29
4	0.6	34	61	-20.28	-11.34	-8.11	91.97
5	4.7	24	54	-27.28	-7.24	-18.11	131.12
6	1.7	65	77	-4.28	-10.24	22.89	-234.39
7	9.4	44	81	-0.28	-2.54	1.89	-4.80
8	10.1	31	93	11.72	-1.84	-11.11	20.44
9	11.6	29	93	11.72	-0.34	-13.11	4.46
10	12.6	58	51	-30.28	0.66	15.89	10.49
11	10.9	37	76	-5.28	-1.04	-5.11	5.31
12	23.1	46	96	14.72	11.16	3.89	43.41
13	23.1	50	77	-4.28	11.16	7.89	88.05
14	21.6	44	93	11.72	9.66	1.89	18.26
15	23.1	56	95	13.72	11.16	13.89	155.01

16	1.9	36	54	-27.28	-10.04	-6.11	61.34
17	26.8	58	168	86.72	14.86	15.89	236.13
18	29.9	51	99	17.72	17.96	8.89	159.66
Sum	215.0	758	1465	-0.04	0.08	0.02	1085.61
4321.02	4321.02	4321.02	4321.02	4321.02	4321.02	4321.02	4321.02
Mean	11.94	42.11	81.28				
Σ(sqrs)	4321.02	35076	131299	12389.61	1752.96	3155.78	

Soil Samples	$X_1 X_2$	$X_1 Y$	$X_2 Y$	x_1y	X_2y
1	0.16	25.6	25.6	199.411	-188.179
2	9.20	24.0	1380	245.571	406.661
3	58.9	220.1	1349	90.875	237.571
4	20.4	36.6	2074	229.975	164.471
5	112.8	253.8	1296	197.507	494.041
6	110.5	130.9	5005	43.827	-97.969
7	413.6	761.4	3564	0.711	-0.529
8	313.1	939.3	2883	-21.565	-130.209
9	336.4	1078.8	2697	-3.985	-153.649
10	730.8	642.6	2958	-19.985	-481.149
11	403.3	828.4	2812	5.491	26.981
12	1062.6	2217.6	4416	164.275	57.261
13	1155	1778.7	3850	-47.765	-33.769
14	950.4	2008.8	4092	113.215	22.151
15	1293.6	2194.5	5320	153.115	190.571
16	68.4	102.6	1944	273.891	166.681
17	1554.4	4502.4	9744	1288.659	1377.981
18	1524.9	2960.1	5049	318.251	157.531
Sum	10139.50	20706.20	63825	3231.48	2216.44

Substituting the values, we have

$$b_1 = \frac{(315578 \times 3231.48) - (1085.61 - 3231.48)}{(1752.96 \times 3155.78) - (1085.6)^2} = 1.7898$$

Similarly, $b_2 = 0.0866$ such that;

$$a = Y - b_1 X_1 - b_2 X = 81.28 - (1.789 \times 8911.94) - (0.0866 \times 42.11) = 56.26$$

and the prediction equation: $Y = a + b_1 X_1 + b_2 X_2$

That is, $Y = 56.26 + 1.7898 X_1 + 0.0866 X_2$

Predicted values

Soil Sample	X_1 (ppm)	X_2 (ppm)	Y	Y_p (ppm)	$Y - Y_p$
1	0.4	53	64	61.6*	2.4*
2	0.4	23	60	59.0	1.0
3	3.1	19	71	63.4	7.6
4	0.6	34	61	60.3	0.7
5	4.7	24	54	66.7	-12.7
6	1.7	65	77	64.9	12.1
7	9.4	44	81	76.9	4.1
8	10.1	31	93	77.0	16.0
9	11.6	29	93	79.6	13.4
10	12.6	58	51	83.8	-32.8
11	10.9	37	76	79.0	-3.0
12	23.1	46	96	101.6	-5.6
13	23.1	50	77	101.9	-24.9
14	21.6	44	93	98.7	-5.7
15	23.1	56	95	102.4	-7.4
16	1.9	36	54	62.8	-8.8
17	26.8	58	168	109.2	58.8
18	29.9	51	99	114.2	-15.2
Sum				1463.0	

*The predicted value Y_p can be estimated for each soil sample from the fitted regression. For soil 1 for example, Y_p = *56.26 + 1.789 8 (0.4) + 0.0866(53) = 61.6ppm*. The predicted value for each soil sample has been calculated this way and tabulated in the table above. The quantity $Y - Y_p$ is the deviation, which measures the failure of the X's to predict Y.

For soil 1, $Y - Y_p = 64 - 61.6 = +2.4$ ppm. The other values have been calculated and tabulated also in the table.

FURTHER ANALYSIS

In some disciplines such as plant breeding genetics and ecology, further analysis beyond the analysis of variance is often necessary and several statistical techniques are available for such further analysis. Some of these techniques include:

(i) Principal component analysis
(ii) Cluster analysis
(iii) Generation mean analysis

FURTHER ANALYSIS OF VARIANCE TECHNIQUES

In agricultural research studies, the main focus is to increase productivity. Productivity may be in terms of yield. That is, increase in quantity, quality, and other desirable attributes such as size and shape. In order to achieve the set objectives, it is imperative to carry out the needed experiments either in glasshouses or in the field with well-considered experimental designs. The general rule is that the simplest design is likely to provide the required precision. A good experimental design is necessary for the collection of data so that differences among individuals or differences associated with the way the data were collected can be removed from experimental error. Some of the designs commonly used in agricultural research are the following:

(i) The Complete Randomized Design
(ii) The Randomized Complete Block Design
(iii) Latin Square Design
(iv) The Split Plot Design
(v) The Split-Split Plot Design

The statistical analysis must reflect the effort in these designs to reduce sampling variation. In this section, only examples of the use of Randomized Complete Block design will be given because it is the most commonly used in field trials.

RANDOMIZED COMPLETE BLOCK DESIGN

This design will be employed to find out the effect of different levels of nitrogen fertilizer on the grain yield of a maize cultivar (FARZ 27).

The nitrogen levels:

N_o - No nitrogen applied
N_1 - 25 KgN/ha
N_2 - 50 KgN/ha
N_3 - 100 KgN/ha

Number of replicates: 5

The grain yield (t/ha) at the harvest can be arranged into Blocks and Treatments as follows for ease of analysis.

	Nitrogen Levels				
Block	N_o	N_1	N_2	N_3	TOTAL
1	5.5	6.4	6.6	7.1	25.6
2	5.7	6.0	6.3	6.8	24.8
3	4.6	5.1	5.6	6.7	22.0
4.	3.2	3.6	4.7	5.3	16.8
5	2.7	3.4	4.0	3.9	14.0
Total	21.7	24.5	27.2	29.8	103.2
Mean	4.34	4.90	5.44	5.96	

Grand Total, G = 103.2 Total number of values (N) 5 X 4 = 20

Analysis:

Step 1: Calculate Correction Factor (CF)

$$CF = \frac{G^2}{N} = \frac{(103.2)^2}{20} = 532.512$$

Step 2: Calculate the Total corrected sum of squares (TCSSQ)

Total SOS = (Total uncorrected sum of squares) - CF

$$= 5.5^2 + 6.4^2 \ldots\ldots 4.0^2 + 3.9^2 - CF = 566.46 - 532.512 = 33.948$$

Step 3: Calculate Block BSSQ

$$\text{Block BSSQ} = \frac{\text{Sum of squares of block total}}{\text{Number of replications per block}} - CF$$

$$= \frac{25.6^2 + 24.8^2 + 22.0^2 + 16.8^2 + 14.0^2}{4} - CF$$

$$= 558.16 - 532.512$$

$$= 25.648$$

Step 4: Calculate Treatment TSSQ

$$\text{Treatment TSSQ} = \frac{\text{Sum of Squares of treatments Total}}{\text{Number of replications per treatment}} - CF$$

$$= \frac{(21.7^2 + 24.5^2 + 27.2^2 + 29.82)}{5} - 532.512$$

$$= 7.292$$

Step 5: Calculate Residual Sum of Squares (RSSQ)

RSSQ = SOS total – SOS blocks – SOS treatments

= 33.948 – 25.648 – 7.292

= 1.008

Step 6: Calculate the degrees of freedom (d-f)

Blocks d.f. = r – 1 = 5 – 1 = 4

Treatment d.f. = t – 1 = 4 – 1 = 3

Residual d.f. = Total d.f. – Block d.f. – Treatment (d.f)

= 19 – 4 – 3 = 12

Step 7: Complete Analysis Of Variance Table

Source	Deg of freedom	SOS	MS	VR
Block	4	25.648	6.4120	76.33
Treatments	3	7.292	2.4307	28.94
Residual	12	1.008	0.0840	
Total	19	3.9483		

NOTE:

(i) Mean square (MS) is calculated by dividing the sum of squares by degree of freedom.

(ii) The variance ratio (VR) is obtained by dividing the treatment mean square (2.4307) by the residual mean square (EMS, 0.0840)

The VR for a block is calculated by dividing the Blocks Mean Square (BMS = 6.4120) by RMS.

Step 8: Conduct F – test

The Variance Ratio (VR) for treatments is 28.94 on (3,12) d.f. To test at 5% level the hypothesis of no difference in treatment effects, we look up ($F_{3\ 12,\ 5\%}$) in the 5% F-table. The entry in column 3 and row 12 is 3.49 i.e. $F_{(3,\ 12,\ 5\%)} = 3.49$. Our value of 28.94 is much greater so we conclude that there is evidence that the treatments have different effects.

For 1% the $F_{(3,\ 12\ 1\%)} = 5.95$ is still greater, so there is strong evidence of a difference in the treatment effects. For 0.1% the value is 10.80. The result is still greater, so there is a very strong evidence of different treatment effects.

In some cases, it may be necessary to quote a probability P-value. In this example, $P<0.001$. This means there is less than 0.1% chance of getting a Variance Ratio as great as the value of 28.94 if there were no underlying treatment differences.

Step 9: Compare treatment means.

Many researchers do not carry out this step if VR for the treatment is not significant. However, if the main aim of the experiment is to compare a particular treatment, say a control treatment with each of others, it is still valid to carry out a Least Significant Difference (LSD) even if VR is not significant.

Step 10: Calculate the Standard Error of the Difference (SED) between two treatments' means. This can be obtained using the RMS value from the analysis of variance table.

$$SED = \frac{\sqrt{2 \times RMS}}{r} = \frac{\sqrt{2 \times 0.084}}{5} = 0.1833$$

Step 11: Find the Least Significant Difference (LSD), start with 5% level.

$$LSD + 5\% = t_{(R.d.f., 2.5\%)} \times SED$$

R.d.f. = residual degree of freedom (12)

$t_{(12,\ 2.5\%)} \times SED = 2.179 \text{ X } 0.1833 = 0.399$ t/ha

Step 12: Compare differences between means with LSD value. In this example, treatment 1 (N_o) is the control treatment so we may wish to compare treatments 2, 3 and 4 with treatment 1.

$$N_1 - N_o = 4.90 - 4.34 = 0.56$$
$$N_2 - N_o = 5.44 - 4.34 = 1.10$$
$$N_3 - N_o = 5.96 - 4.34 = 1.62$$

All these differences are greater than the LSD value of 0.399 so we conclude that the three nitrogen levels are significantly different from the control.

We can go on and find LSD at 1%. At this level, only N_2 and N_3 are significantly different from N_o and at 0.1%, N_2 and N_4 are still significant. The conclusion one can draw here is that the grain yield of maize at 50 KgN/ha is not significantly different from the grain yield of maize grown with 100 KgN/ha.

DUNCAN MULTIPLE RANGE TEST

Treatment means can also be compared using Duncan multiple range test. The test is the most widely used of several multiple range tests available. It gives protection against making mistakes inherent in the indiscriminate use of the LSD test. The test is identical to LSD for adjacent means in an array, but requires progressively lower values for significance between means that are widely separated in the array. This test is used most appropriately when several unrelated treatments are included in an experiment. For example, it is very useful for making all possible comparisons among the yielding abilities of several varieties.

The test involves the calculation of shortest significant differences (SSD) for all possible relative positions between the treatment means when they are arrayed in order of magnitude. The SSDs are then used in an orderly procedure to determine the

statistical difference among the means. The formula SSD is R (LSD) where R is a value from a table of significant studentized factor and is chosen according to the level of significant desired degree of freedom for error and the relative separation of means in the array. For further details see *Statistical Methods in Agricultural Research* by T.M. Little and F.J. Hills.

FACTORIAL EXPERIMENTS

Introduction

In most agricultural research studies, there are always several factors (environmental and treatment) involved. For example, if an Agronomist wishes to assess the economic potential of a new variety of a crop, such as cowpea, he would be required to produce data on the response of the crop to fertilizers, density, insecticides and weeding requirements. To test the variety for each individual requirement would be tedious, time consuming and expensive. In addition it gives no measure of possible interactions between different factors and in statistical terms may give a poor estimate of standard error.

All these problems can be overcome by using a **Factorial Experiment**. Factorial experiments are appropriate when there are two or more types of treatments to be applied in different amounts, which can be applied together or alone. In statistical terms, this sort of treatment is called a ***Factor;*** and the amount applied is the ***Level*** of the factor. The essence of the factorial experiment is that the treatments are made up of all possible combinations of the different factors. For instance in a two-factor experiment there may be

A = NITROGEN FERTILIZER
B = INSECTICIDES

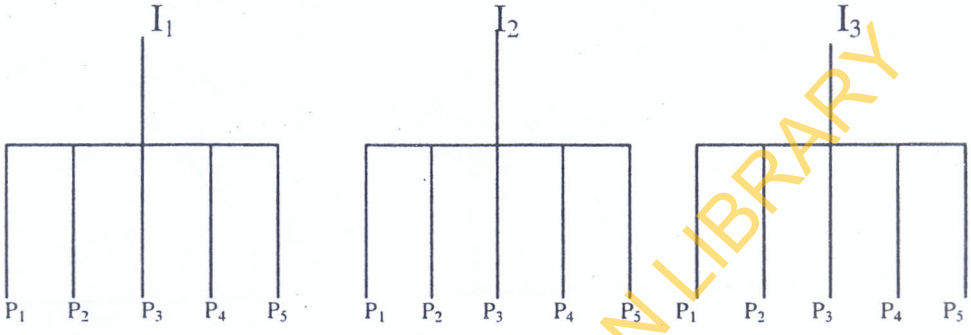
In a three-factor experiment, the additional factor may be

C = SEEDING RATE

Example with interpretation

A factorial experiment was carried out in order to determine the response of cowpea cultivar (Vita 4) to 3 levels of a new

insecticide for the control of Maruca (I_1 I_2 I_3) and 5 levels of phosphate fertilizer (P_1 P_2 P_3 P_4 P_5). The experiment was replicated 3 times. This is a 3 ×5 factorial experiment having the following (3 × 5 ×3) 15 combinations.



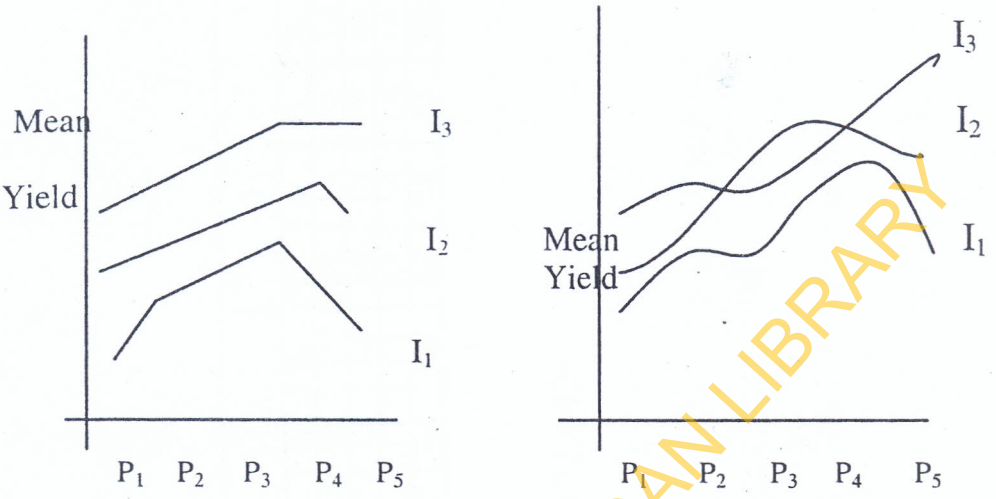
The treatment combinations are:

I_1P_1	I_2P_1	I_3P_1
I_1P_2	I_2P_2	I_3P_2
I_1P_3	I_2P_3	I_3P_3
I_1P_4	I_2P_4	I_3P_4
I_1P_5	I_2P_5	I_3P_5

A randomized design with 3 replications was adopted. The 15 treatments were randomly assigned to the plots within each block. A separate randomization was carried out for each block. This experiment can be used to explore the following objectives.

1. The main effect of phosphorus fertilizer on the yield of cowpea when averaged over levels of insecticides.
2. The main effect of insecticide on the yield of cowpea when averaged over levels of phosphorus.
3. ***Interaction:*** If the response curves to nitrogen fertilizer are not the same for the three levels of insecticides, this is equivalent to saying that the differences between insecticide yields are not the same at each phosphorus level. If the increase in yield in going from one

phosphate fertilizer level to the next is not the same for all insecticides this indicates interaction (see Fig.)



In the first diagram there is no interaction as the response curves are parallel. I_3 gives higher yields than I_2 and I_1 by fixed amounts at each level of P. The second illustrates interaction. I_3 gives higher yields than I_2 and I_1 at phosphorus levels P_1, and P_2 and P_5 but gives lower yields than I_2 at phosphorus levels P_3 and P_4.

If there is a significant interaction it does not make sense to compare the main effects of nitrogen or insecticides. We should compare the effects of insecticide at each level of phosphate or the effects of phosphate at each level of insecticide.

Data For Analysis

BLOCK	I_1					I_2					I_3					
	P_1	P_2	P_3	P_4	P_5	P_1	P_2	P_3	P_4	P_5	P_1	P_2	P_3	P_4	P_5	**Total**
I	0.9	1.2	1.3	1.8	1.1	0.9	1.1	1.3	1.6	1.9	0.9	1.4	1.3	1.4	1.2	**19.3**
II	0.9	1.3	1.5	1.9	1.4	0.8	0.9	1.5	1.3	1.6	1.0	1.2	1.4	1.5	1.1	**19.3**
III	1.0	1.2	1.4	2.1	1.2	0.8	0.9	1.1	1.1	1.5	0.7	1.0	1.4	1.4	1.3	**18.1**
Total	**2.8**	**3.7**	**4.2**	**5.8**	**3.7**	**2.7**	**2.9**	**3.9**	**4.0**	**5.0**	**2.6**	**3.6**	**4.1**	**4.3**	**3.6**	**56.7**

Grand Total, G = 56.7 N = 45

$$\text{Correction factor, } CF = \frac{(56.7)^2}{45} = 71.442$$

$$\text{Uncorrected sum of squares } \sum x^2 = 75.7$$

$$\text{TCSSQ (total)} = 75.73 - CF = 4.288$$

$$\text{BSSQ (BLOCK)} = \frac{\text{Sum of square block total}}{\text{No. of yields per block}} - CF$$

$$= \frac{(19.3^2 + 19.3^2 + 18.1^2)}{15} - CF = 71.506 - 71.442 = 0.064$$

$$\text{TSSQ (treatments)} = \frac{\text{Sum of squares of treatment totals}}{\text{No. of yields per treatment}} - CF$$

$$= \frac{(2.8^2 + 3.7^2 \ldots\ldots\ldots\ldots\ldots 4.3^2 + 3.6^2)}{3} - CF$$

$$= 75.117 - 71.442 = 3.675$$

The Residual Sum of Squares (RSSQ) is found as follows:

RSSQ = TCSSQ (total) – BSSQ (BLOCK) – TSSQ (treatments)

$$= 4.288 - 0.064 - 3.675 = 0.549$$

The treatment TSSQ (SOS (treat) which has 14 d.f can be split into three parts, i.e.

TSSQ (treat) = TSSQ (phosphate) + TSSQ (Insecticide) + TSSQ (Interaction)

To facilitate the calculation of these SOSs see the following table:

	P_1	P_2	P_3	P_4	P_5	**Total**
I_1	2.8	3.7	4.2	5.8	3.7	**20.2**
I_2	2.5	2.9	3.9	4.0	5.0	**18.3**
I_3	2.6	3.6	4.1	4.3	3.6	**18.2**
Total	**7.9**	**10.2**	**12.2**	**14.1**	**12.3**	**56.7**

From this table the following calculations are made:

$$\text{SOS (phosphate)} = \frac{\text{Sum of Squares of Phosphorus}}{\text{No. of yield per phosphate level}} - CF$$

$$= \frac{(7.9^2 + \ldots\ldots\ldots\ldots\ldots + 12.3^2)}{9} - CF$$

$$= 73.932 - 71.442 = 2.490$$

(We divide by 9 because there are 9 original yields for each nitrogen level).

$$\text{TSSQ (Insecticide)} = \frac{\text{Sum of Squares of Insecticides Total}}{\text{No. of yields per insecticide level}} - CF$$

$$= \frac{(20.2^2 + 18.3^2 + 18.2^2)}{15} - CF$$

$$= 71.611 - CF = 0.109$$

TSSQ (Interaction) = TSSQ (treatment) – TSSQ (phosphate) – TSSQ (Insecticide)

$$= 3.675 - 2.490 - 0.169 = 1.016$$

Below is the ANOVA table :

Source	d.f	SSQ	MS	VR	df used
Blocks	2	0.064	0.0320	1.63	
Phosphate (P)	4	2.490	0.6225	31.76	on (4.28 d.f)
Insecticide (I)	2	0.169	0.0845	4.31	(2,28 d.fN X
P x I	8	1.016	0.1270	6.48	(8,28 d.f)
Residual	28	0.549	0.0196		
P x I (interaction)	44	4.288			

Interpretation

From statistical tables ($F_{8\ 28}$, 1%) = 3.23, so the interaction value of 6.48 is highly significant. In order to interpret the treatment effects we should make comparison using a two-way table of phosphate and insecticide means (see table below).

The appropriate standard error is

$$SED_{(\text{int eraction})} = \frac{\sqrt{2 \times RMS}}{r} = \frac{\sqrt{2 \times 0.0196}}{3} = 0.1143$$

where r = number of replications
RMS = Residual Mean Square

If the interaction had not been significant we could compare overall Nitrogen means using:

$$SED_{(P)} = \frac{\sqrt{2 \times RMS}}{r \times g} = \frac{\sqrt{2 \times 0.0196}}{9} = 0.066$$

where r = replications (3) and g = no. of levels of insecticide (3)

Also, the overall insecticide means can be compared:

$$SED_{(I)} = \frac{\sqrt{2 \times RMS}}{r \times n} = \frac{\sqrt{2 \times 0.0196}}{15} = 0.0511$$

where r = replications (3) and n = no. of levels of phosphate fertilizer (5)

Table of Phosphate Fertilizer x Insecticide Means

	P_1	P_2	P_3	P_4	P_5	Mean
I_1	0.93	1.23	1.40	1.93	1.23	1.35
I_2	0.83	0.97	1.30	1.33	1.67	1.22
I_3	0.87	1.20	1.37	1.43	1.20	1.21
Mean	0.88	1.13	1.36	1.57	1.37	

OTHER SPECIAL STATISTICAL ANALYSES

Economic Analysis: Economic Evaluation of Plantation Establishment in Forestry

Introduction

Any productive economic activity produces benefits in the form of goods and services and involves costs in the form of materials consumed and the time of productive factors diverted from other useful employment. According to Worell (1970), a comparison of these benefits and costs gives information for policy decisions. A consideration of benefits and costs leads to a rather obvious basic economic criterion. An activity should not be undertaken unless its total benefits exceed its total costs. An example of economic evaluation of land use deals with production between two alternatives as indicated below.

Economic Evaluation of Land Use

Given the two hypothetical tables below, at a discount rate of 10%, two methods of plantation establishment (Taungya and Direct plantation) were employed on an 8-year rotation to determine economic efficiency of land use. Each method was allotted one hectare of land. As a forest economist, which of these two methods of plantation establishment will be prescribed for economically efficient use of land? State attributes of economic efficiency of your choice.

Objective

To prescribe the plantation establishment method that is economically efficient for land use.

Data Preparation

Table 1: Cash Flow Analysis: *Tectona grandis* Direct Plantation (₦/ha) on eight year rotation in Osun State.

Year	Costs (₦)	Benefits (₦)	Discount Rate 10%	Present Value Costs (PVC)	Present Value Benefits (PVB)	Present Value Benefits minus Present Value Costs (PVB-PVC)
1	650.24	-	0.909	591.07	-	-591.07
2	103.16	-	0.826	85.21	-	-85.21
3	51.76	-	0.751	38.87	-	-38.87
4	19.66	-	0.683	13.43	-	-13.43
5	19.66	-	0.621	12.21	-	-12.21
6	19.66	-	0.564	11.09	-	-11.09
7	19.66	-	0.513	10.09	-	-10.09
8	-	3,850	0.467	-	1797.95	1797.95

Table 2: Cash Flow Analysis: Taungya Teak Plantation (₦/ha) on eight year rotation in Osun State.

Year	Costs (₦)	Benefits (₦)	Discount Rate 10%	Present Value Costs (PVC)	Present Value Benefits (PVB)	Present Value Benefits minus Present Value Costs (PVB-PVC)
1	150.10	3.50	0.909	136.44	3.18	-133.26
2	102.20	7.50	0.826	84.42	6.20	-78.221
3	51.76	10.20	0.751	38.87	7.66	-31.21
4	19.66	-	0.683	13.43	-	-13.43
5	19.66	-	0.621	12.21	-	-12.21
6	19.66	-	0.564	11.09	-	-11.09
7	19.66	-	0.513	10.09	-	-10.09
8	-	3,850	0.467	-	1797.95	1797.95

Economic Analysis

In order to prescribe the plantation establishment that is economically efficient, the following criteria are used for calculation:

(A) *Net Present Value (NPV):* This measures the profit or surplus income from a project after the project has satisfied the rate of return on capital desired by investor. This rate of return desired by investor is used to discount both the costs and revenues of the project. Net Present Value (NPV) is estimated with the formula:

$$NPV = \sum_{t=1}^{n} \frac{(B_t - C_t)}{(1+i)^t}$$

where B_t = Benefits in each project year t
C_t = Costs in each project year t
n = Number of years to the end of project
i = Discount rate

Computation involves six steps

1. A discount rate of 10% obtained from the bank
2. Yearly cost and benefits identified till the rotation age of projects or period under investigation
3. Cost and benefits discounted and presented for every year
4. Costs and benefits estimated in monetary value per year
5. Discounted gross costs and benefits estimated for the rotation year
6. The net values added to obtain NPV

(B) *Benefit-Cost Ratio(B/C)*: The B/C Ratio expresses the sum of the discounted benefit as a ratio of the sum of the discounted cost. According to this criterion, one accepts a project for implementation if the B/C is equal to or greater than one.

Formula:

$$B/C = \frac{\sum_{t=1}^{n} \frac{B_t}{(1+i)^t}}{\sum_{t=1}^{n} \frac{C_t}{(1+i)^t}} \geq 1$$

The same steps above are applicable for computation.

Formula for Discount Rate = $\frac{1}{(1+i)^t}$

Calculation

A. Table 1 (Teak Plantation –Direct)

(i)	Present Value Cost (PVC)	= ₦ 761.97
(ii)	Present Value Benefit (PVB)	= ₦ 1797.95
(iii)	Net Present Value (NPV)	= PVB – PVC =
		= N (1797.95 -761.97)
		= N **1035.98**
(iv)	Benefit- Cost Ratio (B/C)	= ₦ 1797.95 = **2.36**
		₦ 761.97

NPV = ₦ 1035.98 while B/C = 2.36

B. Table 2 (Taungya Teak Plantation)

(i)	Present Value Cost (PVC)	= ₦ 306.55
(ii)	Present Value Benefit (PVB)	= ₦ 1814.99
(iii)	Net Present Value (NPV)	= PVB – PVC
		=₦ (1814.99 – 306.55)
		= N **1508.44**
(iv)	Benefit- Cost Ratio (B/C)	= ₦ 1814.99 = **5.92**
		₦ 306.55

NPV = ₦ 1508.44 while B/C = 5.92

Interpretation and Inference

For economically efficient use of land employing 1 hectare for two methods of plantation establishment each, **Tungya Teak plantation** is prescribed because the NPV (**₦1508.44**) and corresponding B/C (**5.92**) are greater than Direct Teak NPV (**₦ 1035.98**) and B/C (**2.36**). Taungya is more economical, desirable, acceptable and profitable than Direct Teak plantation.

Attributes of Economic Efficiency

i. Discounted benefit of Taungya is greater than discounted benefit of Direct Teak plantation
ii. Discounted cost of Taungya is less than Discounted cost of Direct Teak plantation

iii. Net Present Value of Taungya is greater than Net Present Value of Direct Teak Plantation
iv. Taungya is used for conflict resolution between government and rural dwellers where there is land hunger
v. Taungya can compete in the short run with agriculture for land use.

REFERENCES

Clewer, A.G and Scarisbrick, D.H (1991) *An introduction to the Principle Crop Experimentation,* BASF publication ISBN 0 86266 1900

Worrel, A. C. 1970. *Principles of Forest Policy*. New York: McGraw-Hill.

CONTENTS