KENYA FORESTRY RESEARCH INSTITUTE

TRAINING NOTES

ON

STATISTICAL METHODS

AND

COMPUTER APPLICATIONS

IN

BIOLOGICAL SCIENCES

Compiled by James E. Chiria

KEFRI TRAINING WORKSHOP IN BIOMETRICS
27TH – 29TH August 2002
K.E.F.R.I. Conference Hall

COMMONWEALTH
SECRETARIAT

Kenya Forestry Research Institute

Acknowledgement

These notes for SERIES A of the KEFRI Training Workshop in Biometrics have been prepared to assist the participating KEFRI scientists and those from collaborating institutions, follow the training sessions. It is hoped that the participants will thereafter use them often as reference in the application of the concepts in their investigation. The notes will also form the foundation for SERIES B, which will emphasise on applications in on-farm trials, survey methodology and complex experimental designs.

Dr Alice A. Kaudia, Assistant Director Service Program, supervised preparations leading to SERIES A of Training Workshop. During the period, very useful comments were received from Dr B. Kigomo, Dr Oballa, Dr Chagala and Dr Ongugo on the course content and where emphasis should be placed.

Richard Coe and Peter Muraya of ICRAF, John Knowles and Sonal Nagda of ILRI, J. Ngugi-Mwangi of KARI and Prof. J. Odhiambo, Head of the Department of Mathematics, University of Nairobi, gave further comments to enrich the training materials and focus.

PROGRAM FOR SERIES A OF TRAINING WORKSHOP

			Time			
Date	Session	Topic	From	To	Duration	Resource Persons
Tue 27th		Opening by the Director	9:00:00 AM	9:15:00 AM	0:15:00	Director KEFRI, Dr. P. Konuche
"	1	***Statistics and the scientific method of research***	9:15:00 AM	10:15:00 AM	1:00:00	James Chiria, KEFRI
"		Tea-break	10:15:00 AM	10:30:00 AM	0:15:00	
"	2	***Measures of Location and Spread***	10:30:00 AM	11:30:00 AM	1:00:00	Jason Kariuki/James Chiria
"		Break	11:30:00 AM	11:40:00 AM	0:10:00	
"	3	***Practicals exercises on sessions 1 & 2***	11:40:00 AM	12:40:00 PM	1:00:00	Resource Persons
"		Lunch break	12:40:00 PM	2:15:00 PM	1:35:00	
"	4	***SPSS session for measures of location and spread***	2:15:00 PM	3:45:00 PM	1:30:00	Resource Persons
"		Tea-break	3:45:00 PM	4:00:00 PM	0:15:00	
"	5	***Distribution theory***	4:00:00 PM	5:00:00 PM	1:00:00	Thomas Achia, UON
Wed 28th	6	***Basic concepts in Analysis of Variance***	8:40:00 AM	10:00:00 AM	1:20:00	Ngugi Mwangi, KARI
"		Tea-break	10:00:00 AM	10:20:00 AM	0:20:00	
"	7	***Oveview of experimental designs***	10:20:00 AM	11:40:00 AM	1:20:00	James Chiria, KEFRI
"		Break	11:40:00 AM	11:50:00 AM	0:10:00	
"	8	***Further concepts in Design & Analysis of experiments & surveys***	11:50:00 AM	1:00:00 PM	1:10:00	Ngugi Mwangi, KARI
"		Lunch break	1:00:00 PM	2:15:00 PM	1:15:00	
"	9	***GENSTAT application to ANOVA (Practicals)***	2:15:00 PM	3:45:00 PM	1:30:00	Resource Persons
"		Tea-break	3:45:00 PM	4:00:00 PM	0:15:00	
"	10	***Comparison of SPSS and GENSTAT in ANOVA (Practicals)***	4:00:00 PM	5:00:00 PM	1:00:00	Resource Persons
Thur 29th	11	***Data Management - Basic data format; the model statement***	8:30:00 AM	10:00:00 AM	1:30:00	James Chiria, KEFRI
"		Tea-break	10:00:00 AM	10:20:00 AM	0:20:00	
"	12	***Data Management- Data verification, analysis, storage and retrievel***	10:20:00 AM	11:40:00 AM	1:20:00	Thomas Achia, UON
"		Break	11:40:00 AM	11:50:00 AM	0:10:00	
"	13	***Practical session on Data Management***	11:50:00 AM	1:00:00 PM	1:10:00	Resource Persons
"		Lunch break	1:00:00 PM	2:15:00 PM	1:15:00	
"	14	***Practical session on Data Management (continued)***	2:15:00 PM	3:45:00 PM	1:30:00	Resource Persons
"		Tea-break	3:45:00 PM	4:00:00 PM	0:15:00	
"	15	Wrap-up session: Discussion and Appraisal of SERIES A	4:00:00 PM	4:45:00 PM	0:45:00	Resource Persons
"		Close by Deputy Director, Dr B. Kigomo	4:45:00 PM	5:00:00 PM	0:15:00	Deputy Director, Dr. Kigomo

Session 1 Tuesday, 27 August 2002 ; 9:15-10:15 am by *James E. Chiria*

STATISTICS AND THE SCIENTIFIC METHOD OF RESEARCH

SUMMARY

This session will deal with:

- *Data collection methods*
- *Concepts of probability*
- *Assumptions made in carrying out statistical tests*
    - *Independence of events*
    - *Normality of observations and residuals*
    - *Equality of variances*

The practical implications of the failures of the assumptions in forestry research will be demonstrated in the theory and the computer sessions.

Introduction

This session will deal with some characteristics of the scientific method and statistical techniques in forestry research. It is often not possible to make observations on all the units or subjects in the population of study. Hence, one resorts to taking a sample, which is hopefully representative of the whole and then one makes inferences. The science of statistics enables one to:

- Select an objective sample
- Make valid generalisations and
- Assign degree of uncertainity in the conclusions.

Data collection methods

Experiments Unlike surveys, experiments involve a degree of interference on nature in the sense that some control is exercised on the units of study. Randomisation is performed with certain restrictions where for example in a split-plot design, the main plots are first randomly assigned within each replicate, then the treatments are randomly allocated to the plots in the main plots in turn.

Nevertheless, objectivity is ensured through random allocation. Moreover, reliability of the conclusions is enhanced through replication or repetition, while the effect of *extraneous factors* on the units is reduced through local control or blocking.

Examples of experiments in forestry are found in silvicultural trials in plantations and nurseries as well as in laboratory trials. In these cases, pre-identified treatments are applied to well-defined experimental units.

Surveys consist of studies in which the observations made do not interfere on nature. Through random sampling, the units of study are selected with *known* probabilities. Surveys can be undertaken for purposes of:

- Estimation of population parameters
- Comparing different populations
- Distribution pattern of organisms
- Finding the inter-relationship among variables and
- Studying how man and his environment interface or relate.

Surveys find great applicability in ecology and wildlife biology.

Simulations as a way of data collection have gained importance in recent times with the advent of computers. These techniques allow experiments to be conducted on the computer through *what-if* studies rather than with real life system. A forest stand as an eco-system is made up of components or elements such as the tree, birds and wildlife.

Through simulations, large-scale field experiments, which are costly and time-consuming. The starting point in simulations is to develop a mathematical model that captures the relevant features of the system.

The elements of the system have certain characteristics or attributes, which can be given numerical or logical values. These elements by interacting among themselves have certain relationships that can be studied through mathematical equations. Thus, the state of the system under alternative conditions can be studied and predicted.

The concept of probability

Event is an occurrence such as getting a red flower or round seeds in a breeding trial; it is usually denoted by E. The events are required to be *mutually exclusive* and *exhaustive* (explanation).

Example Suppose the colour of flowers in a particular plant species is governed by the presence of a dominat gene A in a single gene locus. Suppose further that the genetic combinations with the resulting colours are shown in the following table:

Table 1.1. Genetic combinations of dominant gene A

	A	a
A	red	red
a	red	white

Consider the event of getting red flowers in a progeny through selfing of a heterozygote Aa. With the four possibilities of the gametic combinations AA, Aa, aA and aa assumed to be equally likely, we say the event of getting red flowers occurs with probability ¾ while that of getting white flowers is ¼. This idea leads to the so-called classical definition of probability.

Classical definition Suppose an *event* E can occur in x out of a total of n possible equally likely ways or outcomes. Then the probability of occurrence of the event is given by x divided by n and denoted as:

$$P(E) = x/n$$

Relative frequency definition A major drawback of the classical definition of probability is the vagueness of the words "equally likely". Hence, another approach is to have the estimated or empirical probability of an event is taken as the proportion of the times the event occurs when a large number of observations are made.

Example. In the search of a particular endangered species, we could keep a count of the number of plants examined n and the number of endangered species found, x as more counts are made as illustrated in the table and the figure. As n becomes large, the ratio x/n or the relative frequency will approach a certain limit i.e. the ratio will stabilise about a certain value.

Table 1.2. R elative frequency of occurence of a rare species

Number of plants examined, n	5	10	15	20	30	50	80	120	150
Number of endangered species observed, x	0	1	1	1	2	3	5	7	9
Proportion x/n	0.000	0.100	0.067	0.050	0.067	0.060	0.063	0.058	0.060

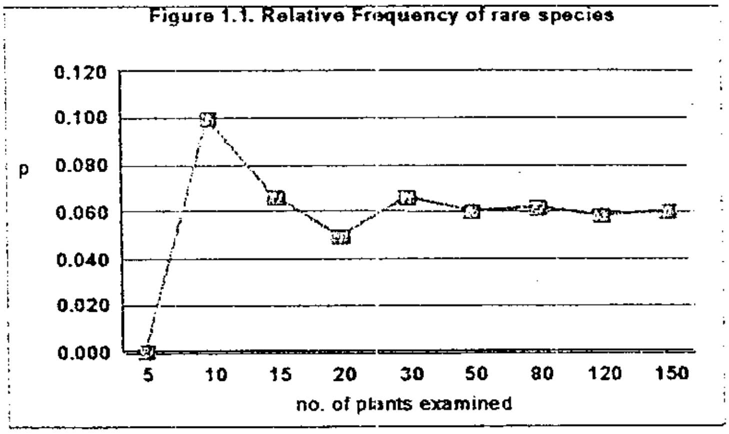


Figure 1.1. Relative Frequency of rare species

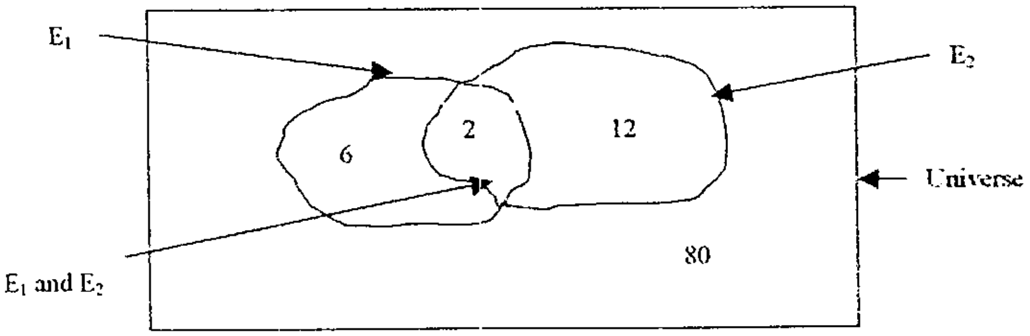
Joint probability relates to the probability of two or more events occurring simultaneously such as getting red flowers and rounded seeds. Here, the key word is *and*; we write P(E_1 and E_2) to denote joint probability.

Conditional probability Suppose E_1 occurs first then followed by E_2. We call the probability of E_2 occurring after E_1 has occurred a conditional probability and this is written as:

$$P(E_2/E_1) = P(E_1 \text{ and } E_2)/P(E_1)$$

These concepts can be represented through the following Venn diagram:

Figure 1.2. Venn diagram showing union of events



The total number of occurrences, n =100. The probability of E_1, $P(E_1) = 8/100$; that of E_2, $P(E_2) = 14/100$ and the joint probability, $P(E_1 \text{ and } E_2) = 2/100$. The conditional probability of E_2 occurring given that E_1 has occurred, is:

$$P(E_2/E_1) = P(E_1 \text{ and } E_2)/P(E_1) = (2/100)/(8/100) = 0.25$$

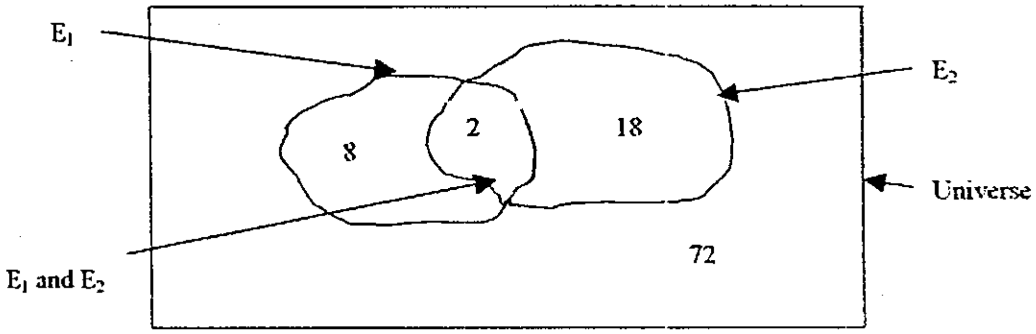
Assumptions made when carrying out statistical tests

<u>Independence of events</u> Suppose as before, E_2 occurs after E_1. If the probability that E_2 occurs is not affected by E_1 having occurred, we say that the two events E_1 and E_2 are independent and we write:

$$P(E_2/E_1) = P(E_2).$$

We can illustrate this through the following Venn diagram:

Figure 1.3. Venn diagram to illustrate independence of events



Here we have $P(E_1) = 10/100$; $P(E_2) = 20/100$ and $P(E_1 \text{ and } E_2) = 2/100$. This time, the conditional probability of E_2 given E_1 has occurred is:

$$P(E_2/E_1) = P(E_1 \text{ and } E_2)/P(E_1) = (2/100)/(10/100) = 0.20$$

which is the same as the as $P(E_2) = 0.20$. Thus, with independent events, the probability with which one event occurs in the universe is unaffected by the occurrence of another another.

Note: (1) In the previous Venn diagram, the chance of E_2 occurring after E_1 is higher (0.25) than that of E2 occurring in the whole population (0.14); we say the two events are linked. (2) In the last case, the probability of E_1 occurring in the universe is 0.10; also unaffected by E_2 occurring first.

Mathematical consequences of independant events

From the equations above, we can see that, if two events are independent, then the probability of them occurring together i.e. their joint probability, is merely a product or multiplication of their separate probabilities, i.e.

$$P(E_1 \text{ and } E_2) = P(E_1)\,P(E_2)$$

In the Venn diagram demonstrating independence, we have:

$$P(E_1 \text{ and } E_2) = 0.02 = 0.1 \times 0.2 = P(E_1)\,P(E_2)$$

Example Consider the joint segregation of the flower colour and the shape of seed in a plant species. Suppose these characters of colour and shape are individually governed by the presence of the dominant genes A and B respectively. Suppose further that the individual combinations and their outcomes are as follows:

	A	a
A	red	red
a	red	white

	B	b
B	round	round
b	round	wrinkled

Table 1.2. Segregation of flower colour and shape of seed

Probality (red flower) =3/4; Probality (wrinked seed) = 3/4

If in the progeny obtained through selfing of a heterozygote AaBb we have the events: E_1 – getting plants with red flowers and E_2 – getting plants with round seeds which we assume are independent events, then the probability of getting plants with red flowers and round seeds in the selfed progeny is:

$$P(E_1 \text{ and } E_2) = P(E_1)\,P(E_2) = \tfrac{3}{4} \times \tfrac{3}{4} = 9/16$$

In the example of segregation considered above, if an experiment is carried out and the proportion of plants with red flowers and round seeds turns out to be far apart from 9/16, we will begin to question the assumption of independence. We then begin to think of whether there is an *interaction* between the two gene loci. The question of how far apart will be answered through *statistical tests of significance.*

A more relevant application of these concepts in agro-forestry could be given by work carried out by Dr Kaudia (1996). One of the objectives was to see whether the type of species of trees grown by the farmers in the study area varied with social status, land ownership or size of farm. The question would be: "Is the probability that a farmer grows *S. sesban* the same for each social status"?

Normality of observations

Another important assumption that will be made in our analyses will be that of normality of observations i.e. they are assumed to come from the normal distribution. If this indeed is not the case, then one of the things we have to do is to transform the data.

Imagine that insect counts (x) made in a certain PSP at three different times of the day are 1, 10 and 100. The square root of these counts y = square root (x) are 1.0, 3.2 and 10.0 whereas the logarithms of the counts z = log (x) are 0, 1 and 2. We see that the logs of the counts are closer together than the their square roots which in turn are closer than the original counts. We say that the logarithmic transformation is more powerful than the square root transformation. When regression work is done, the assumption is that the observations are normal. If you plot a histogram of the data and find that they are skewed (asymmetric), this would call for a transformation to satisfy the assumption of normality. In other tests, it will be required that the residuals also look normal for a valid test.

Consider a set of data[1] collected from a study that followed the activities of *C. capitana* over a 60-day period in the laboratory. The trends in the number of live flies against the days after first emergence and the plots of residuals are shown Figures 1.4-1.6. It is evident that the plot of residuals from the log fit is closer to normality. Examples of other situations where log fits are analysed are in the technical bulletin *"Sampling Cypress Aphids" (1993).*

[1] Unpublished data by P. Nemeye, ICIPE

Fig 1.4 Exponential fit to number of live flies

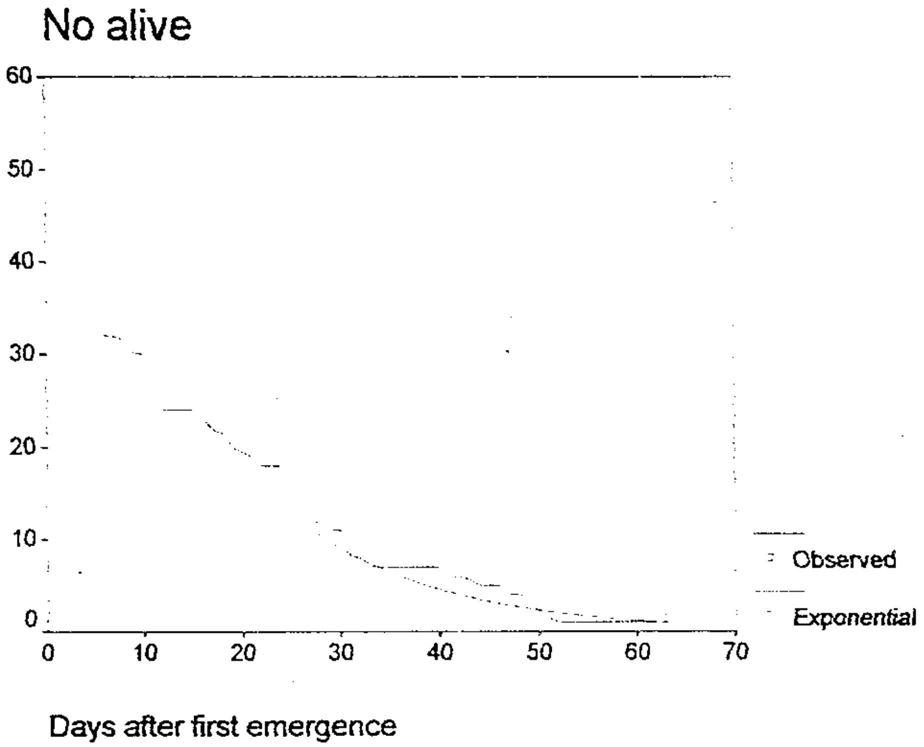


Fig 1.5 Logarithmic fit to number of live flies

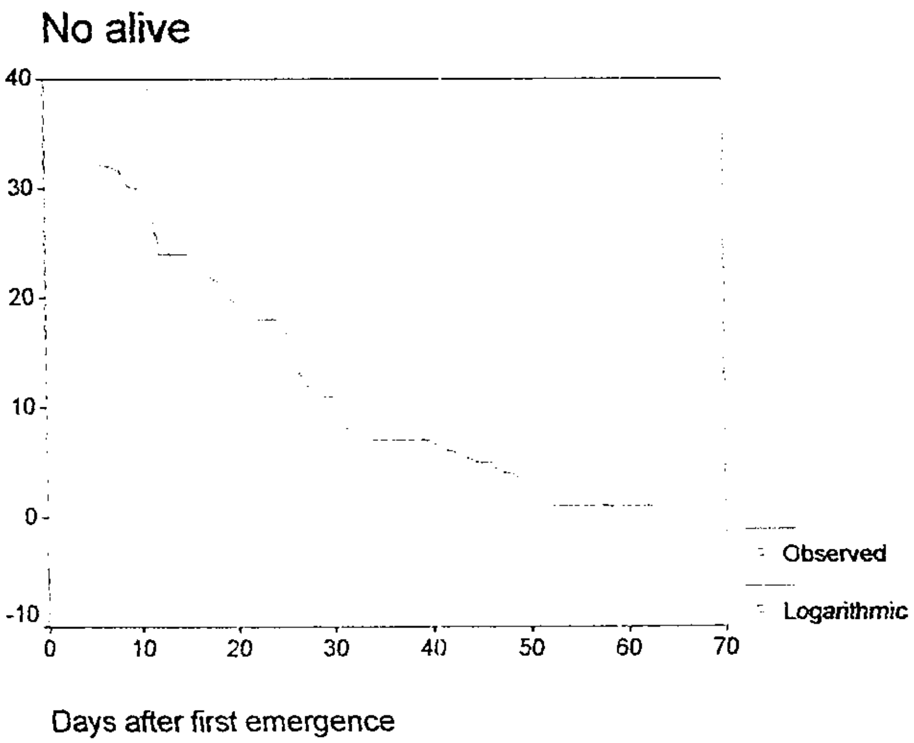


Fig. 1.6 **Histogram of residuals from exponetial fit**

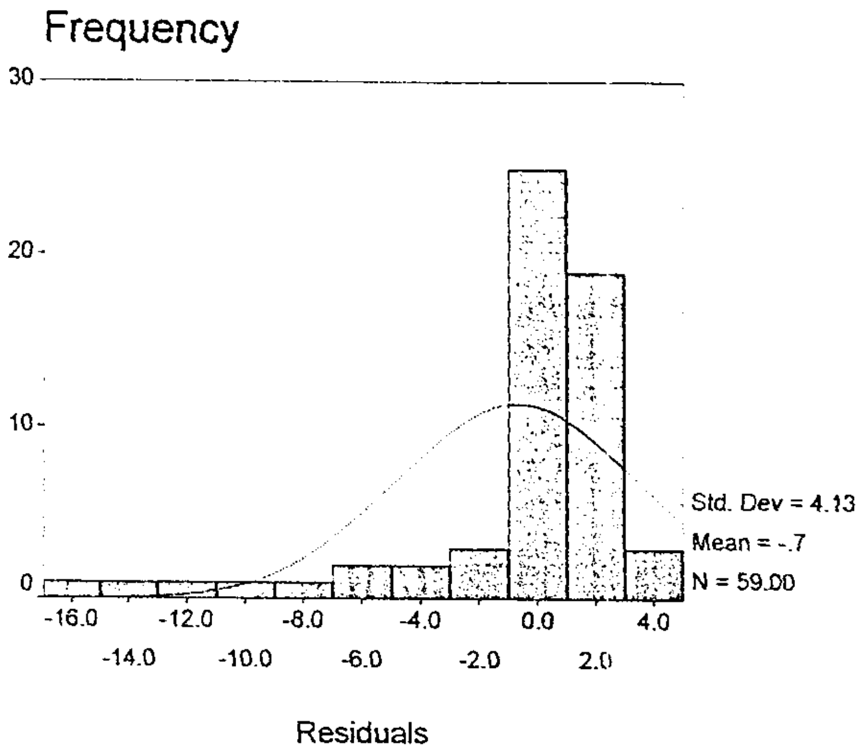
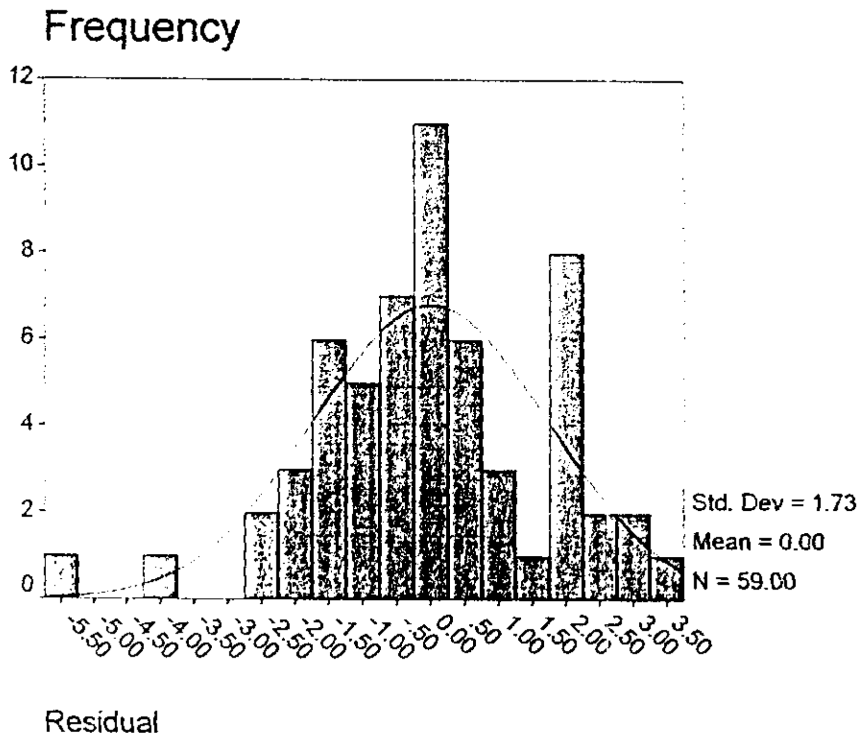


Fig. 1.7 **Histogram of residuals from log fit**



A drawback with transformations such as this is that the researcher may be more comfortable with the original observations rather than the transformed variable.

Equality of Variance

Yet another assumption that is frequently made is that the observations come from distributions that have *equal* variances and yet the reality may be that they are unequal as shown in Figure 1.7

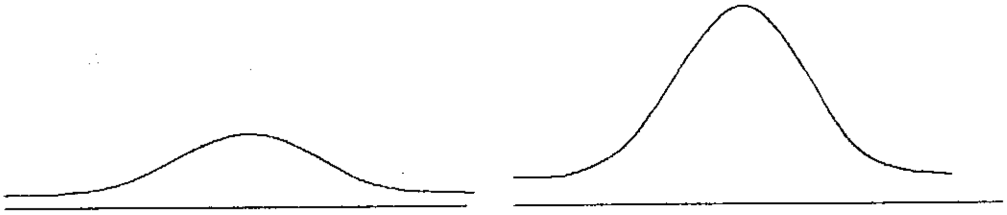


Figure 1.7 Distributions with unequal variances

Practical significance

The conclusions that one comes up with will depend on the validity of the assumptions made. So one should always stop to look back and ask whether the assumptions made were valid. As we have seen, transformations can be performed and also, there are tests that can be carried out to examine these assumptions. Fortunately, some of the tests, such as that of equality of variance have been shown to be *robust* (Box, 1978) meaning that small to moderate deviations from the assumption can be tolerated.

Independence of events will be more difficult to ensure or assume. There are tests such as *Tukey's* one degree of freedom test for non-additivity. Certain situations that one will have to be wary about the need for independence are:

1. Linear regression being used when a time series analysis may be more appropriate because of the dependent nature of the observations;
2. Repeated measures analysed as independent replicates;
3. Within-plot samples as if they were replicates;
4. Complex designs analysed as simple designs and
5. Clustered samples analysed as simple random samples.

Session 2. Tuesday, 27 August 2002, 10:30-11:30am by *J Kariuki and J Chiria*

AVERAGES AND SPREAD OF DATA

SUMMARY

In this session we shall look at:

- *Scales of measurement:*
    - *Nominal; Ordinal; Interval and Categorical*
- *Summarisation of data through:*
    - *Frequency distributions and frequency polygons*
- *Representative values of a set of data (averages):*
    - *The arithmetic mean; the weighted mean; the geometric mean; the harmonic mean; the median and the mode*
- *The spread or variation in the data:*
    - *Range, Standard deviation, Standard error*

The relevance of each of the measures in forestry research will be illustrated. Practical exercises in manual calculations and computer applications will follow to cement the concepts.

Measurement scales

Raw data consist of measurements of some attribute on a collection of individuals and they can be made in one of several scales:

Nominal which is a number or symbol used to classify an object, a person or characteristics e.g. gender or the state of health whether diseased or healthy. Nominal variables provide a list of choices with no meaningful order to the list. The arithmetic mean of a nominal value is useless. Instead, use the mode and run frequencies. You can then run cross-tabulation using the nominal variable. To display the data use pie and bar charts.

Ordinal in which the groups or classes of individuals exhibit pair-wise hierarchy e.g. socio-economic status: low, medium or high, an opinion on a particular issue could be coded 1 (strongly agree) or 2 (somewhat agree) but by how much is not known i.e. they have an implied order between the response choices. When it comes to analysis, we examine the median and mode for these variables and run cross-tabulations or some true-based approaches. Here, the bar charts are good for displaying the choices.

<u>Interval</u> or continuous variables have an implied order and implied distance between the response options e.g. temperature scale. Also, age in years has a one-unit difference that is the same throughout the distribution. Such variables lend themselves to a much wider range of powerful statistics than the previous variables. Regression is one of the more popular statistical procedures using interval variables. Scatter plots and histograms are appropriate graphical displays for these kinds of variables (summary).

<u>Categorical</u> If necessary a continuous variable can be collapsed into a <u>categorical</u> variable whose response options are categories– either nominal or ordinal. However, there are instances where it is appropriate to use a continuous variable but record in a range (Exhibit B A.2). Questions to do with respondent's income in currency increments being sensitive and difficult question to answer are dealt with in this manner to reduce non-response.

Summarisation of data

Regardless of the scale of measurements data can be summarised by distributing it into classes or categories. The number of individuals in each class, called the class frequency can be determined. The tabular arrangement of the data by classes together with the corresponding class frequencies is called a frequency distribution as shown in Table 2.1.

Class	Interval	Class mark	Frequency
1	Less than 100	50	3
2	100 – 300	150	10
3	300 – 500	400	15
4	500 – 1000	750	52
5	1000 - 2000	1500	15
Total			95

Table[2] 2.1. Frequency distribution of volume m^3 of species from Western Kenya

[2] From Dr Muchiri, MCC

The frequency distribution can be graphed by a histogram that consists of a set of rectangles having bases on a horizontal axis. The centres are at the class marks and the lengths are equal to the class interval sizes with the areas being proportional to class frequencies as shown in Figures 2.1 (a) and (b).

A frequency polygon is a line graph of class frequency plotted against class mark obtained by connecting the midpoints of the tops of the rectangles in the histogram as shown in Figure 2.1 (c). However, a point most often overlooked is that if the x-axis is a continuous variable as volume is here, the appropriate plot should be as shown in Figure 2.1 (d).

Further condensation of data is possible through measures of location (averages), dispersion (spread), skewness (departure from symmetry) and kurtosis (degree of peakedness) of the distribution. Of these, we shall only look at the first two in some detail.

Averages

There are many types of averages or Measures of Central Tendency. The common ones are the arithmetic mean, the mode, the median, the harmonic mean and the geometric mean. Each of these has its application the biological sciences.

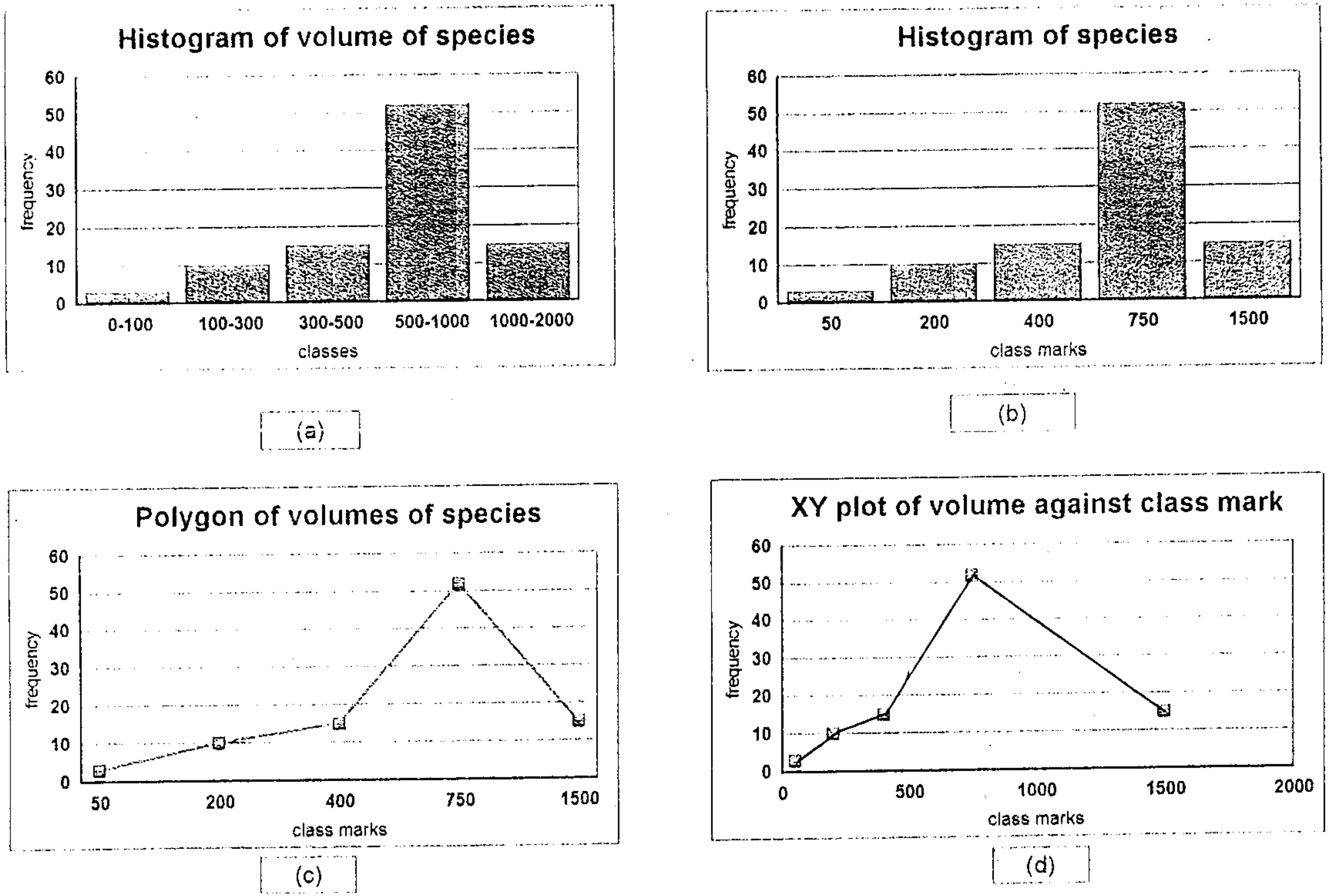
When there is no ambiguity or when no confusion will arise, the mean will be used to refer to the arithmetic mean. This is the most common of the means.

Illustrations. Suppose you have 6 measures of DBH:3, 4, 4, 5 , 6 and 10.

The arithmetic mean is the sum of 3+4+4+5+6+10 = 26; which when divided by 6 gives 6.3.

The median is the value that splits the distribution into exactly two equal parts; the lower and the upper halves. In this example, it is 4.5. When the distribution is split into four (4) equal parts, we refer to quartiles; if into ten (10) parts, deciles; if one hundred parts percentiles etc.

Figure 2.1 Histograms, Polygon and XY Plot of Volume of Species



The mode is the value that occurs in the distribution with the largest frequency. In this example, it is 4.0. A disadvantage of the mode is that it may not be unique.

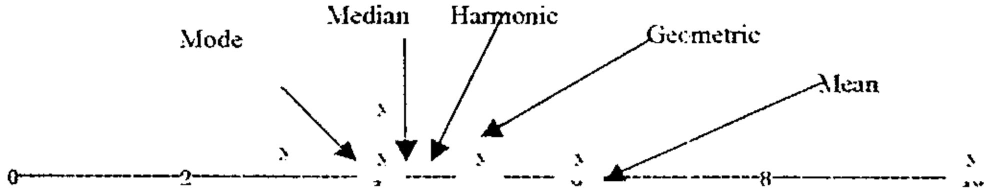
The geometric mean is used to describe the *average* of data that exhibit growth over time such as in population dynamics (bacteria growing in a petri-dish). In our example, to compute it, we first multiply the six observations to obtain their *product*: 3x4x4x5x6x10= 14,400. We then take the sixth root of the product and obtain 4.9.

The harmonic mean is appropriate when describing the representative value of rates, ratios or proportions. We first obtain the *reciprocal* of a number; this is one (1) divided by the number: the reciprocal of 10 is 0.1, of 2 is 0.5, of 4 is 0.25 etc. The sum of the reciprocals is obtained. In our case, this is:

1/3 + ¼ + ¼+1/5 + 1/6 +1/10 = 1.3

The number of observation, here six (6) is then divided by the sum of reciprocals i.e. 6/1.3 to obtain 4.6.

The averages we have calculated can be represented on a linear scale as follows; in which it is seen that a representative or average value of data collected depends on the nature of the study! The same set of data can have different averages.



x – original observations

We shall see that for the normal distribution, the (arithmetic) mean, the median and the mode are the same; otherwise, the distribution will be said to be skewed or slanted.

The mean is used most because, among other characteristics, it occurs frequently in real life and it has some nice mathematical properties. A disadvantage of the mean is that it is affected by extreme values that pull it to one side. This can be illustrated by the following example. Let us suppose that the last value was wrongly recorded as 100. Then we clearly see that the mean is the average value most affected.

The weighted mean   Consider the case where the DBH of three species have been measured as follows:

Parameter	*C. Lusitanica*	*E. Saligna*	*P. Patula*	Total
Total	90	54	171	315
Count	3	2	4	9
Mean	30.00	27.00	42.75	35.00

Table 2.2. Derivation of weighted mean

The overall mean (35.00) where sample sizes (3, 2, 4) are not the same, is called a weighted mean. It is different from the unweighted mean or simple average of 30.00, 27.00, 42.75 which is 33.25. The weighting comes about because for each mean we must take into account the number of observations on which the mean is based. Thus we have

$$3 \text{ x } 30.00 + 2 \text{ x } 27.00 + 4 \text{ x } 42.75 = 315$$

This total is divided by the sum of the weight: 3 + 2 + 4 = 9, to give 35.00. It is clear that this is merely the total of the diameter measurements divided by the total number of observations. It is worth noting that when all sample sizes are the same, then the unweighted average or mean will be the same as the weighted mean. In general, weighting will be important to take into account.

Spread in data

Apart from the average or representative value, we also need to know how far apart the data are or by how much do they vary about each other and also about their *average* value. (Two locations with same mean temperature but different SD). There are some measures: the range, semi-inter-quartile range, and the standard deviation. In the example above the range, the difference between the maximum and minimum observation made is 10-3=7.

For the standard deviation, we first calculate the deviation of each observation from the their (arithmetic) mean (3.5) is shown in column (2).

Note that the sum of the deviations about the (arithmetic) mean is zero (.0). It can be shown that if you calculate the deviations about any other average and sum them, you do not get zero as you do in the case of the arithmetic mean. We refer to this as a nice mathematical property of the arithmetic mean. This makes the mean an important statistic to use.

i	1	2	3	4	5	6	Sum	Mean
x	3	4	4	5	6	10	32	5.3
$x - \bar{x}$	-2.33	-1.33	-1.33	-0.33	0.67	4.67	0.00	-
$(x - \bar{x})^2$	5.4444	1.7778	1.7778	0.11111	0.4444	21.7778	31.3333	6.2667

Table 2.3. Derivation of standard deviation

The sum of the squares (SS) of these deviations is 31.3333. When this sum is divided by the degrees of freedom (6-1=5), we get 6.2667. (DF is important in small sample statistics). The square root of this number is the standard deviation, which is 2.50.

Following the comments made in the introduction to this notes, we shall then say that the set of data {3,4,4,5, 6 and 10} is summarized by the two measures: 5.3 and 2.50. These measures are called statistics, which are linear combinations of original observations. Alternatively, we say the data set is now represented in the range 6.3 plus/minus 2.50 or lie in the range 3.8 – 8.8.

The inverse weighting Suppose you have two estimates of a parameter with their corresponding sample variances as shown in the table.

i	1	2	sum
estimate, e	10.0	16.0	26.0
sample variance	2	4	6
inverse weights, w	0.50	0.25	0.75
e x w	5.0	4.0	9.0

Table 2.4. Derivation of inverse weighted mean

For the combined estimate, one average would be the ordinary arithmetic mean, 26/2 = 13.0. But taking into account the spread in the data, the estimate from sample 1 has a higher precision than that from sample 2. So it makes sense to expect the combined estimate to be closer to 10.0. This is achieved by taking the inverse of the sample variance in which sample 1 has a higher weight (0.50) than sample 2 (0.25). We multiply the estimate by the inverse weight in the last row and sum, which is then divided by the sum of the weights to obtain 9.0/0.75 = 12.0. In the practicals we shall see the effect of having sample 2 estimated with a higher precision.

Session 3 and 4. Practical exercises and computer applications

Tuesday 27 August 2002, 11:40-12:40; Lunch break; 2:15-3:45 pm
Resource Persons

Exercises to demonstrate transformations using SPSS and GENSTAT

Example. Suppose we have the following set of nine (9) measurements of DBH:

{28, 30, 32, 26, 28, 39, 41, 45, 46}

Verify that the following statistics are obtained:

arithmetic mean = 35.0, mode = 28.0, median = 32,
geometric mean = 34.25, harmonic mean = 33.52.

These can be represented on a linear scale together with the original values as follows. Notice the relative positions and you can imagine the pros and cons of each MCT.

For the spread in the data, the range is 18 whereas the standard deviation is 7.8 so that the data can be said be said to lie in the range: 27.2 – 42.8. [This is one reason why extreme values (7) need to be looked at more closely when checking or analysing data].

Other examples to be given to show how SPSS, GENSTAT summarise data.

Question For the example of inverse weighting given in the lecture, complete the table with the sample variances as shown.

i	1	2	sum
estimate, e	10.0	16.0	xxx
sample variance	4	2	xxx
inverse weights, w	xxx	xxx	xxx
e x w	xxx	xxx	xxx

Compare the estimates obtained with equal weights (1,1) and using the incorrect weights (4,2).

Session 5. Tuesday, 27 August 2002; 4:00-5:00 Thomas Achia, UoN

THE DISTRIBUTIONS

SUMMARY

In this session we shall look at the following theoretical distributions:

- *The Binomial Distribution*
- *The Multinomial Distribution*
- *The Poisson Distribution*
- *The Normal distribution (N)*
- *The Student's, Distribution (t)*
- *The Chi-squared Distribution*
- *The F- Distribution*

We shall stress where these distributions apply in forestry research.

Introduction

Probability distributions can be thought of as the theoretical or ideal limiting forms of the relative frequency distributions when the class intervals are made smaller and the number of observations become large. In this sense then, probability distributions can be thought of as being distributions for populations whereas relative frequency distribution can be thought of as distributions of samples that are drawn from this infinite or hypothetical population.

The Binomial distribution

The Binomial Distribution arises in a situation where in a single trial you obtain a "success" or "failure". This is an either/or, yes/no situation. For example, an entomologist might observe the number of seedlings affected by termites in a plantation where 25 seedlings were planted in each plot. Each plant in the plot is either affected (success) or not affected (failure). The number of plants affected by termites in a plot is said to follow a binomial distribution.

An experiment was carried out in Kakuzi to study the effect of Fipronil 3G on control of termites. The number of surviving trees out of 25 originally planted per plot and the attacked trees is given in the table below:

Plot no.	1	2	3	4	5	6	Total
Number surviving	23	24	22	22	20	14	125
Number attacked	8	10	12	22	20	8	80

Table 5.1 Frequency of terminate attack in Kakuzi experiment on effect of Fipronil 3G

The rate of attack on a single tree is estimated by $p = 80/125 = 0.64$. If we now took another random plot of 25 trees and assumed the termites' infestation follows a binomial distribution, the probability of getting x trees attacked out of the 25 is given by the formula:

$$P(\text{x trees attacked}) = {}^{25}C_x\,(0.64)^0\,(0.36)^{n-x}$$

A couple of the probabilities are shown in: We would then say that with a general infestation rate of 64%, the chances of getting no tree attacked out of 25 is very low. The average number of infested trees would be 0.64x25 = 16.0 with a standard deviation of 2.4.

The Multinomial distribution

The multinomial distribution is a natural extension of the binomial distribution in that it looks at the joint distribution in more than two categories. In the Kakuzi example, we would be looking at the trees falling in several disease categories: ...

The Poisson distribution

The Poisson distribution is asymmetrical i.e. skew. The distribution is characterised by the parameter lambda, which measures the average number (of organisms say), which are found in unit area of space. It turns out that the variance of the number of organisms is also lambda.

The Poisson distribution applies for example, in ecological studies where certain sparsely occurring organisms are randomly distributed over space. In forestry, this applies to the random distribution of trees or plants in natural forest as opposed to plantations. In such circumstances, observations on the number of organisms found in small sampling units are found to follow the

Poisson distribution. Other areas of application are e.g. purity or germination counts in seed testing, insect counts, and bacterial colonies on plates.

Example An ecologist observes the number of seedlings of a rare species found in 100 quadrants taken from a block of natural forest. Suppose the total number of plants of the species found in the 100 quadrants is 10. Then the mean number of plants per quadrant is 10/100 = 0.1 with standard deviation $(0.1)^{0.5} = 0.32$.

Assuming that the number of plants per quadrate follows a Poisson distribution, the probability of getting zero plants in a fresh quadrate from the same forest tract is.

$$P(\text{no plants obtained}) = (0.1)^0 \exp(-0.1)/0! = 0.9048$$

and the probability of obtaining at least one plant is 1-0.9048 = 0.0952, which clearly shows the nature of the rare event.

The Normal Distribution

The binomial and Poisson distributions are for discrete variables that take the values 0,1,2, ... up to infinity. The normal distribution is one where the variable is a continuous random variable. It is a theoretical distribution of immense practicability. It was originally proposed as a measurement error model but has found to be the basis of variation in a large number of biometrical characters. It is characterised by the parameters mu (average) and sigma (standard deviation).

If the deviations of the original observations from the mean, $(y - \mu)$ are divided by the standard deviation, σ we obtain the standardized normal score z and we write:

$$z = (y - \mu)/\sigma$$

When observations are standadised this way, the re-scored variables theoretically now take on all values between negative and positive infinity; they are centred at zero and have unit standard deviation.

This enables all distributions to be 'standardised' so that reference is made to one and only one table.

The Distribution of sample means

If we repeat an experiment several times, we would generate a series of means. A valid question to ask would be how well does a single mean represent the (true) population mean?

Imagine going into a population and each time taking a random sample of fixed size, n and you do this with replacement i.e. after recording the measurements, say yield, you put the unit or individual back into the population. Suppose further that you kept on doing this repeatedly, each time calculating and recording the means. If now you constructed a frequency distribution of the means, you would obtain what is illustrated in figure 1.

With increasing sample size (n), the distribution of means becomes *narrower* and *taller*. The standard deviation becomes smaller but the mean remains the same. It is because of this (theoretical) relationship that we can claim to be able to estimate the variance of the population mean from a single sample.

The standard deviation of a single observation, σ_Y is called the standard error per unit or plot. The standard deviation of the population of means, σ/sqrt(n) is called the standard error of the mean or merely the standard error. The distribution of means can also be standardized as:

$$z = (\bar{y} - \mu)/\sigma_{\bar{y}}$$

The Student's t-Distribution

Often, the standard deviation of the population is not known but we can estimate it from a small sample of data. In 1908, a chemist by the name W. S. Gosset, who wrote under the pseudonym "Student" he assumed sampling from a population with known mean but unknown variance. Each time a sample of size n is drawn with replacement from such a population, he calculated the sample mean and it's standard error. He found that the ratio of the deviation of the sample mean from the population mean when divided by the standard error:

$$t=(\bar{y}-\mu)/s_{\bar{y}}$$

followed a distribution he called the Student's t-distribution with n-1 degrees of freedom (DF). The assumptions that are made for deriving the t-distribution are:

1. The original observations are normally distributed about a population with known mean (μ) but unknown variance (σ^2) and

2. The sample standard deviation s is <u>independent</u> of the observations, y

The precise form of the t-distribution depends on the degree of uncertainty in the sample variance, which is measured by the number of DF (or equivalently) the sample size. When the DF is infinite, there is no uncertainty in the estimate of the sample variance. Hence the t-distribution becomes the standard normal distribution, z as shown in Fig. 2.

However, when the number of degrees of freedom is small, the possibility of variation in the sample variance result in a greater probability of extreme deviation and hence in a "heavier" tailed distribution. It is to be noted that except in the extreme tails of the distribution, the normal distribution provides a fair approximation to the t-distribution when the degrees of freedom is at least 15. (Ref BHH).

The Chi-square Distribution

The chi-square arises out of the distribution of the sum of squares that can never be negative. Whereas the normal and the t-distributions take all negative and positive values, the chi-square distribution takes only positive values. In ANOVA, the mean squares follow the chi-square distribution. It is a continuous distribution.

However, the chi-square is often used in the distribution of samples based on counts that are discrete or discontinuous variables. The interest is usually to see whether variables are related or are independent. It is also to test whether a group of samples could have been drawn from the same population. Another area of application is to see whether observed and expected values are close. The difference between observed and expected values is squared and the squares are summed after dividing by the expected values. In all cases of discrete variables the Chi-square tables only give us approximate probabilities of obtaining deviations from expectations, as large or larger than those calculated on the basis of chance alone.

Cross tabulations are appropriate when you have two or more categorical variables. Continuous variables do not lend themselves to cross tabulations since you would get as many rows or columns as there are different responses. When looking at cross tabulations, if the probability of a chi-square is 0.05 or less, it usually small enough for one to feel that the distribution did not result from chance but that the patterns that are seen are 'real'. The chi-square statistic does not measure the *strength* of the relationship but rather it measures if the relationship is due to chance. A significant a chi-square would lead you to study the crosstabulation and characterize the pattern that you have seen. (Mutitu data)

The F-Distribution

This ratio was called F, by George W. Snedecor in honour of Sir Ronald A. Fisher, a pioneer in the use of mathematical statistics in agriculture. The F test is a ratio between two variances and is used to determine whether two independent estimates of variances can be assumed to be estimates of the same variance. In ANOVA, the F test is used to answer the question; can it be assumed that treatment means resulted from sampling populations with equal means (and hence equal variances)?

Session 6. Wednesday, 28 August 2002; 8:30-10:00 am Ngungi-Mwangi K.A.RI.

OVERVIEW OF EXPERIMENTS

SUMMARY

In this session we shall look at:

- *Scientific research as an iterative process of learning*
- *Need for careful and precise definition of objectives*
- *How objectives lead to identification of treatments*
- *Incorporation of non-statistical knowledge*
- *Random variation in experiments*
- *Replication for precision of estimates*
- *Choice of site, layout and blocking*

Introduction

For a given tree species such as *C Lusitanica*, if we have measured the diameter at breast height (DBH) we can compute the basal area through a formula. This would be deduction. If we had no idea of the relationship between the variables: height (h), diameter (d) and volume (v), we would chop up the economic portion of the tree into logs; measure the height and diameter of each log separately. For each log, we would measure the volume by say the displacement method and then add these up for the whole tree. We would then establish the relationship between h, d and v and come up with say, Sterling's formula. This would be induction.

Thus with deductive reasoning, you move from the general to particular whereas with induction you start off with a particular set of observations from which you want to generalize. Scientific research is a process of guided learning, which can be depicted as a feedback loop shown in Figure 1 (Box, 1978). The object of statistical method is to make that process as efficient as possible.

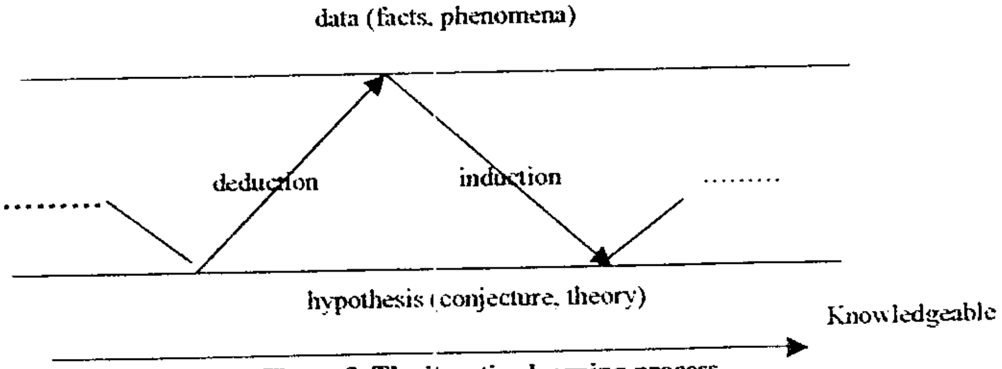


Figure 3. The iterative learning process

In the iterative process of learning, you have a feeling that such and such a species/provenance does well and therefore that it should do equally well in a given environment (tentative hypothesis). You deduce the consequences of such a hypothesis by collecting data. In another situation, a biotechnologist inoculates seeds and feels that the resultant plants should have certain desirable nutritious values for animal fodder. Again, this leads to deductions resulting in observations being made and data collected. When the consequences of the hypothesis and the data collected fail to agree, this discrepancy can lead to the modification of the hypothesis and we have the next deduction-induction. As time goes on, we become more knowledgable.

Objectives of the experiment

In the recent past, participatory on-farm experimentation has become increasingly important in KEFRI's research activities. Coe (1998) reviewed the experiences of research involved in this area together with the role of biometrics. One of the things he noted was that most of the problems could have been avoided if: "The *research objectives* had been clearly stated at the start and *conflicting objectives* not addressed in the same research activity" (emphasis author's).

It is extremely important to clearly define the objectives of the study. When this is done as a team, all team members must agree on the objectives. The team must also agree on the criteria to be used in judging that the stated objectives have been achieved. Equally important, should the objectives change, then all team members have to be made aware of this. They must then all agree on the revised objectives together with the new criteria.

A team of scientists involved in an entomology study may have wanted to look at the behavior of a certain insect species with regards to their water and food intake. Suppose now there is need to also look at the amount of time spent on other activities (walking, resting, pheromone calling, mating, ovipositing) and that sex differences are important, then this must be communicated to the team and the members must all agree on the new objectives and how their success will be judged.

Treatments

Treatment is a generic term borrowed from agricultural research that stands for fertilizers, methods, processes, materials or whatever the things are that are being compared. Usually the objectives of the investigation would lead directly to the treatments to be applied as well as the measurements that will be made and the analyses to be performed.

The treatments can be single factor such as when four tree species are being compared or when the three rates say of P applied at 0, 50 and 100 kg/ha are being compared. In this case, we have one factor; species at 4 levels while the single fertilizer has 3 levels.

We could have more than one factor each at different levels. For example, an experiment under consideration to be carried out by a KEFRI scientist is to see effect on germination of seed under 4 different land preparation methods in each of which seed processed in 3 different ways are to be planted. This leads to a factorial arrangement designated as 4x3. Where in addition, he might want each seed preparation to be either planted in the soil or broadcast, we would now have a third factor, planting method at 2 levels. The arrangement is now a 4x3x2 factorial.

The need for control treatments would depend on the objective of the experiment. Consider the cases where the objective were to assess the gain from adding fertilizer, then a 'no fertiliser' plot would be necessary as control. On the other hand if the objective were to compare organic and inorganic fertilizer, there would be no need for a 'no fertiliser' plot. Other situations will be discussed in the practical sessions.

Use of nonstatistical knowledge

A meeting was held recently between the scientists from the Farm Forestry Unit and the KEFRI biometricians to discuss the design of an on-farm *calliandra* trial to test the performance of inoculants. The suggestion was made that on terrace farming, the control plot should always be on the top. The biometrician asked why the position should not be randomized as for other treatments. The explanation was given that the microbes move especially when it rains. This would mean that one would not know whether the yield in the (lower) control plot in is not because of the movement of the microbes from the top plot. Moreover, the objective of the trial is to convince the farmer to use the inoculated seeds. The biometricians were educated.

This anecdote underscores the importance of the relationship between scientific investigation and statistics. "It is possible for a scientist to conduct an experiment or investigation without statistics (but) it is impossible for a statistician to do so without scientific knowledge. *However, a good scientist becomes a much better one if he uses statistical methods*" (Box, 1978). The intellect and knowledge of the investigator will be put to best use when effective statistical tools are at his disposal but the more important thing is to be able to design an experiment efficiently. There are also certain situations where randomization may not be possible.

Chances occurrence or random variation in nature

Nature behaves in a complex manner but by observing the results of experiments, we try to predict its behaviour. One reason why experiments are conducted is because in nature, there is chance occurrence (random fluctuations), which we are not able to control. This fact can be appreciated when we note that even when experiments are repeated, they will not yield identical results.

It is because of the chance variations in nature that no two plants will exhibit the same growth performance even when they are of one species/provenance. So there will always be random variation to contend with and which we must always keep in mind when conducting experiments. *We can reduce or control random variation but we can never eliminate it.* One could philosophise that it is good that there is random variation; otherwise it would be difficult to judge the success of an experiment!

Relevance of chance variations to experiments

A good experimental design then will be one where the random variation is reduced to as small a value as possible. Thus, we should account *as much as possible* for variation in data due to factors whose effects we are studying and leave *as little as possible* to chance variation. The significance of treatment effects (to be defined) will depend on how much bigger the variation due to treatments is compared to random variation; the bigger the variation due to effects compared to random variation, the more significant the effect(s) will be. If the experiment has not accounted well for effect of treatment, the data can be thought of as being '*noisy*' and that it has not given any '*signals*'.

Randomisation

Another requirement is to have the treatments affected equally, whether positively or negatively, by chance variation. This is done through **randomization.** By this procedure, we ensure that we are not biased when allocating treatments to experimental units or plots of land. In surveys, we want to select units to be included (or excluded from the study) with an equal chance.

The mechanics of how this is carried out will be covered during the practicals session but suffice it to say that from any random start in the random number tables that you can generate different random sequences. But also, these days computer software exist for automatically the random numbers.

Replication

Replication is repeating of the <u>same</u> treatment to <u>more than one</u> experimental <u>unit</u> or plot. Often replication is confused with <u>blocking</u> but technically they are different and they serve different purposes.

When discussing design of experiments, we shall see that with the complete randomized block design (CRBD), a block coincides with a replicate but that as a general guideline, *blocks are contained in replicates*. The other cautionary note is that when you are sub-sampling from an experimental plot you are **NOT** replicating in the way we have defined replication of a treatment.

Replication serves several purposes:

- Intuitively if a treatment has been applied only once, this is not appealing
- Through repeats, we have a better *average* of treatment performance or effect;
- You have some insurance should something going wrong with some plots (cases abound of this) and
- The investigator is able to widen range of validity of results.

Choosing the site and blocking

If your experimental site slopes in one direction, this fact must be taken into account. You would allocate your treatments in fairly homogenous portions or blocks of land (illustrate with diagram). Soil variability is generally the largest likely source of variation but there is also topography and other physical features to be considered. It is important to account for these sources of variation.

Additionally, where resources permit, a uniformity trial may have to be carried out in the first season and/or take soil samples to help identify patterns in soil variability. A map of the site and information about its past history come in handy. It may well turn out that parts of the site may have to be omitted such as where there are anthills. Hence a visit to the *proposed* site by the team of investigators and the biometrician(s) is extremely important.

In KTRI, an experiment has been set up to study the effect of tree shade on yield of tea. The trees are aligned in an E-W direction to take into account the sun's radiation. You may have to take into account the direction of wind in relation to the surrounding tree plantations. These are biophyiscal conditions that must be factored into the design.

A rule of thumb to help decide when to block and randomize is: "*Block what you can and randomize what you can't*".

Other issues

Since the purpose of SERIES A is to review concepts participants have come across in earlier courses in statistics and have a synopsis of the major issues, there are certain important details that will be deferred to the next series. These include size, shape, guard, borders and orientation of the plots, estimating values for missing plots, management issues among others.

SESSION 7   Wednesday 28 August 2002; 10:20-11:40; J. E. Chiria

BASIC CONCEPTS OF ANALYSIS OF VARIANCE

SUMMARY

In this session we shall see how to decide whether two or more samples come from the same population by:

- *Stating the null hypothesis;*
- *Assigning a level of rejection or significance;*
- *Defining treatment effect and calculating residuals;*
- *Partitioning sums of squares and setting up ANOVA table;*
- *Comparing treatment means and*
- *Giving an insight into statistical computer analysis.*

The stress put at the end will be that even if these principles are illustrated for the completely randomized design, the ideas are easily extended to other complex designs.

Introduction

We have seen that raw data collected in a single sample from a certain population; can be summarized through its average and spread. If you have two or more samples, you might want to see whether they come from the same population. This is achieved through analysis of variance (ANOVA), a technique developed by R. A. Fisher through which total variability is apportioned into its various sources.

The Null Hypothesis

Suppose in comparing the growth performance of two species: *Eucalyptus grandis* (EG) and *Robusta gravillus* (RG), you make ten measurements of the DBH of trees of each species. You want to know whether, on the basis of diameter, the species exhibit the same growth. In carrying out the test, our starting point is to assume that there is **no** difference in the mean diameters of the two species. This is the so-called **null hypothesis**, which in symbols is put as:

$$H_0: \mu_1 = \mu_2$$

This statement is construed to mean that the two samples come from the **same** population. Some important characteristics of the null hypothesis are that it must be *neutral*, *unambiguous*, *exact* and *non-controversial*. However, to every null hypothesis, there are several **alternative** hypotheses such as:

$$H_1: \mu_1 < \mu_2; H_2: \mu_1 \leq \mu_2; H_3: \mu_1 > \mu_2; H_4: \mu_1 \geq \mu_2; H_5: \mu_1 \# \mu_2$$

i.e. the samples come from populations whose means differ in several relative directions. Our task comes down to: When do we **reject** the null hypothesis?

Courtroom analogy

The judge in a court case may on the basis of the evidence heard (the data collected), send an accused person to jail or set that person free. We do not really know whether the accused is guilty or innocent. Similarly in our statistical test, we shall be led to either <u>reject</u> the null hypothesis (imprison an innocent person) or <u>not reject</u> it (let go free a guilty person).

It is desirable to have a judicial system whereby as few innocent people as possible are sent to prison (***Type I*** error). We would also want as few guilty persons as possible to be let loose (***Type II*** error). The following table illustrates this.

Table 3. Type of error by decision made and true state of affairs

DECISION	TRUE STATE OF AFFAIRS	
	Innocent	Guilty
Imprison (Reject)	***Type I***	O.K.
Set free (Not reject)	O.K.	***Type II***

[*Type I* is the more serious error to commit: we want a maximum of 5 innocent persons out of 100 accused to be sent to jail. *Type II* is an <u>opportunity cost</u>: you will tolerate let go free a maximum of 5 guilty persons out of 100 accused].

And when we do not reject the null hypothesis, we do so *tentatively* thus leaving room for the possibility that in the future, we might come out with evidence leading us to reject it.

Level of significance

When we reject the null hypothesis, we shall assign ourselves a maximum amount of risk that we want to take. This limit is called the **level of significance** (usually 5%, 1% 0.1%) of our test.

Example A KEFRI scientist taking forest inventory in Western Kenya measured the DBH of three species as shown in the table. The (weighted) mean of the diameters is 35.00.

Table 4. Diameters (cm) of trees by species

Case	*C. Lusitanica*	*E. Saligna*	*P. Patula*	**Total**
1	28	26	39	-
2	30	28	41	-
3	32	-	45	-
4	-	-	46	-
Total	**90**	**54**	**171**	**315**
Count	**3**	**2**	**4**	**9**
Mean	**30.00**	**27.00**	**42.75**	**35.00**
Species effect	**(5.00)**	**(8.00)**	**+7.75**	*(5.25)*
Note: - N.A.; Weighted mean = 35.00; Unweighted mean = 33.25				

We would want to know whether on the basis of the diameter measurements the species could be claimed to come from the same population. Let us define a few terms.

The Treatment (seedlot, species etc) effect

The treatment effect is defined as the difference of the treatment mean from the overall (weighted) mean. For the three species the estimates of these effects are shown in the second last row of Table 4. Note that the sum of effects in the last column is –5.25 but that when the weights (counts) are taken into account, the sum is 0 i.e.

$$3 \times (-5.00) + 2 \times (-8.00) + 4 \times (+7.75) = 0$$

We shall see that when the sample sizes are the same, the sum of the effects will be zero. Consequently, as in this example, we only need to know 2 of the effects and the third can be deduced; hence we have two (2) degrees of freedom for species.

Residuals (estimates of error)

Each observation made deviates from the general mean by a certain amount. If we further take away treatment (species) effect, we are left with *residuals*, which give us a measure of the chance variation. Thus, each observation is broken down into: the mean effect, species effect and residual. All this is illustrated in Table 3.

Sum of Squares

Next, we calculate the squares of: the observations, the effects and the residuals and sum these as shown in Table 4. The total sum of squares of the original observations is referred to as the raw or uncorrected sum of squares (RSS), which in the example is 11,511. The sum of squares due to the mean, often called the correction factor (CF) is here 11, 025. The sum of squares of the species effects or the treatment sum of square (TrSS) is 443.25. And finally, the error sum of squares (ESS) is 42.75.

Recap

Here we are looking at data that have arisen from a completely randomized design (CRD). So far, we have seen how we can decompose each observation into the various effects and compute estimates of error. This breakdown will be done according to a *model* we have in mind that we *tentatively* entertain. We will need to further analyse the residuals for any patterns that might make us revisit the model for improvement. We have also seen how the sum of squares of the effects and that of the residuals add up to the total or raw sum of squares. When we next look at randomized complete block design (RCBD) (and indeed other more complicated designs), exactly the same principles will apply. With RCBD, we shall introduce the block effect but everything else will be the same. Later on we shall introduce the concepts of *linear* and *additive* effects.

Table 3. Decomposition of observations into mean and species effects and residuals

Observed			=	Mean			+	Species			+	Residuals		
28.00	26.00	39.00		35.00	35.00	35.00		(5.00)	(8.00)	7.75		(2.00)	(1.00)	(3.75)
30.00	28.00	41.00		35.00	35.00	35.00		(5.00)	(8.00)	7.75		0.00	1.00	(1.75)
32.00	-	45.00		35.00	-	35.00		(5.00)	-	7.75		2.00	-	2.25
-	-	46.00		-	-	35.00		-	-	7.75		-	-	3.25
315.00				315.00				0.00				0.00		

Table 4. Decomposition of raw sum of squares into the mean, species and residuals sum of squares

Observed			=	Mean			+	Species			+	Residuals		
784	676	1521		1225	1225	1225		25.00	64.00	60.0625		4.00	1.00	14.0625
900	784	1681		1225	1225	1225		25.00	64.00	60.0625		0.00	1.00	3.0625
1024	-	2025		1225	-	1225		25.00	-	60.0625		4.00	-	5.0625
-	-	2116		-	-	1225		-	-	60.0625			--	10.5625
SS -> 11511				11025				443.2500				42.7500		
DF -> 9				1				2				6		

Table 5. Analysis of Variance (ANOVA) Table

Source of Variation	*DF*	*Sum of Squares*
Total or Raw or Uncorrected (RSS)	9	11,511.00
SS due to the mean or correction factor (CF)	1	11,025.00
Treatments (or species here) SS, TrSS	2	443.25
Error SS (ESS)	6	42.75

Table 6. Decomposition of the deviations from the mean into and species effects and residuals

Deviations from mean			=	Species			+	Residuals		
(7.00)	(9.00)	+4.00		(5.00)	(8.00)	7.75		(2.00)	(1.00)	(3.75)
(5.00)	(7.00)	+6.00		(5.00)	(8.00)	7.75		0.00	1.00	(1.75)
(3.00)	-	+10.00		(5.00)	-	7.75		2.00	-	2.25
-	-	+11.00		-	-	7.75		-	-	3.25
0.00				0.00				0.00		

Table 7. Decomposition of raw sum of squares of the deviations into the species and residuals sum of squares

Deviations			=	Species			+	Residuals		
49.00	81.00	16.00		25.00	64.00	60.0625		4.00	1.00	14.0625
25.00	49.00	36.00		25.00	64.00	60.0625		0.00	1.00	3.0625
9.00	-	100.00		25.00	-	60.0625		4.00	-	5.0625
-	-	121.00		-	-	60.0625			--	10.5625
SS -> 486.00				443.2500				42.7500		
DF -> 8				2				6		

Table 8. The characteristics of the ANOVA table

Source	*DF*	*SS*	*MS*	*VR (Fc)*	*Ft*	*p-level*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Corrected total SS (CSS)	8	486.00	60.750	-		
Treatments SS, TrSS	2	443.25	221.625	31.11	5.14	0.001
Error SS (ESS)	6	42.75	7.125			

Degrees of Freedom

We made a total of 9 observations and we say that the original observation have 9 degree of freedom (DF).. This is because after making 8 observations there is nothing stopping us making other observation. However, when we calculate the mean (35.00 in this case) we have removed one degree of freedom, and we are left with 8 DFs. Equivalently, when we calculate deviations of the observations from the mean, we only need to calculate 8 deviations; the last one is automatic.We have seen already that for the species, the sum of their effects is zero; hence we have 2 degrees of freedom for treatments. We are then left with six (6) DFs for the residuals i.e. 9 - 1 - 2 = 6.

Analysis of Variance (ANOVA) table

We put all this information in an analysis of variance (ANOVA) table as shown in Table 5. Often, the ANOVA is presented with the SS due to the mean removed to leave the corrected SS for the total (CSS) as depicted in Tables 6, 7 and 8.

The outputs of the ANOVA table will vary depending on what statistical package one is using but with the essentials explained in Table 8. Columns (1), (2) and (3) have already been explained. Column (4) is called, the mean square (MS) and is column (3) divided by column (2); this gives a measure of the 'average' of sum of squares of each of the terms in the model. Column (5) is variously referred to as, the variance ratio (VR) as in GENSTAT, or the calculated F-ratio (Fc) is the mean square of the source of variation divided by the *appropriate* error mean square for that term.

In column (6), Fc is compared to the tabulated F-value (Ft) to test whether effects are significant i.e. whether the null hypothesis should be rejected. Column (7) gives the actual probability or significance level; GENSTAT calls it *F pr* while SPSS refers to it as *Sig.* In some books it is called the probability or **p-level**. Columns (6) and (7) are complementary; the latter gives you the actual probability that such extreme values will occur by chance alone if the null hypothesis is true. It is in this sense, more informative.

The F-test leads us to reject the null hypothesis of no different difference and we conclude that there is enough evidence to say that at least one pair of differences is different. The F-test is an **omnibus** or overall test. We must now proceed to finer details to find out which means are different. [In the more technical sense, we now proceed to find out which *treatment effects* are different]. The species means are presented on a linear scale as shown in Figure 4.

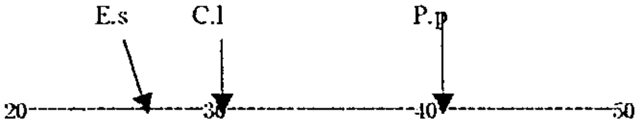


Figure 4. Relative species mean DBH

Table of means and their differences

Table 9. Table of species means and their differences

		C. lusitanica	E. saligna
		30.00 (3)	27.00 (2)
P. patula	42.75 (4)	12.75**	15.75**
C. lusitanica	30.00 (3)	-	3.00 n.s.

With three treatments (A, B and C), there are three mean differences to compare: AB, AC and BC; with 4 treatments, there are six differences etc. From Figure 2, we see that P. patula had a higher mean DBH than the other species. Is this due to chance or is this 'real'? One test that we now carry out is to compute the *least significant difference* (LSD). If the observed difference is bigger than the LSD, we conclude that the means are significantly different. We calculate this for the difference between *P. Patula* and *C. lusitanica* and the other two will be left for the practicals session.

The estimate of error variance of DBH, s^2 from the experiment is 7.125 from the error mean square. The sample variance of the difference between the means of *P. Patula* and *C. lusitanica* is obtained by multiplying s2 by the sum of reciprocals of the sample sizes i.e. 7.125(1/4 + 1/3) = 4.16. The square root of this number is the *standard error* of the difference, (sed) between the means i.e. sed = square root of 4.16 = 2.04. The tabulated Student's –t value for 6 DF at p-level of 0.001 is $t_6(0.001) = 3.71$. The LSD is obtained by multiplying the sed by the tabulated t-value i.e.

$$LSD = sed \times t_6(0.001) = 2.04 \times 3.71 = 7.57$$

For this pair of means, the observed difference is 12.75, which is bigger than the LSD for the pair. Therefore we conclude that *P. Patula* mean diameter is bigger than that of *C. lusitanica*. We continue with the other two pairs similarly. It will turn out that the pair of means of *E. saligna* and *C. lusitanica* are not significant. For scientific publications, the reporting is usually done by saying that means with same letters are not significant as shown in the following table.

Table 10. Usual presentation of means in scientific papers

C. lusitanica	*P.patula*	*E. saligna*
30.00^b	42.75^a	27.00^b

Means with same letters are not significant

How does the computer handle analysis?

Once we have appreciated the mechanics of the steps in ANOVA it becomes easy to see how the computer performs the analyses. Let us recast the data presented in Table 11 in the following row x column structure.

Extension to other designs

The preceding analysis has been one of the so-called completely randomized design (CRD). It is the simplest of designs and assumes no blocking has been done. Among its advantages is the fact that if for some reason, an observation is lost or missing, one does not have to estimate for it. As we go higher and higher to other designs, the restrictions become more and the layout more complex. The next lecture gives an overview of the designs and when they need to be applied. The basic principles introduced here of decomposing the observations into the effects stipulated in the model, the breakdown of the sum of squares into the various components will apply.

Table 10: Adapted from Table 3 of KEFRI ANNUAL REPORT AND RECORD OF RESEARCH July 1999 - June 2000

Effect of inoculation method, using strain KCC 17 on modulation and of *C. calothyrsus* seedlings after 3 months of growth in Leonard jars

	Number of	Nodule dry	Seedling	Shoot	Root dry	Total dry
Method of Inoculation	nodules/plant	wt (g/plant)	height (cm)	wt (g/plant)	wt (g/plant)	wt (g/plant)
KM1: pre-treated seeds prior to sowing in germination trays	89b	0.037a	13.9bc	0.312b	0.101a	0.444b
KM2: pre-treated seeds immediately after sowing in germination trays	59a	0.031a	14.9cd	0.290b	0.088a	0.393b
KM3: after 50% seedling emergence	67a	0.034a	17.1de	0.343b	0.097a	0.482b
KM4: Immediately after transplanting	54a	0.028a	11.5b	0.263b	0.089a	0.381b
KM5: control; uninoculated	NAc	NA	7.8a	0.136a	0.085a	0.227a

Note: a - describes the methods of inoculation

b - Values in a column followed by the same letter are not significant according to *Neumann* and *Keul's* test at $P<0.05$

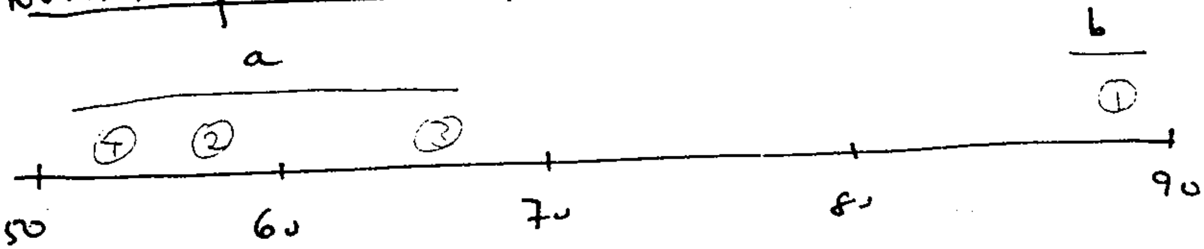
c - Not applicable

Most agro-forestry practices rely on the use of nitrogen-fixing trees and shrubs, but there is still little remarkably little quantitative information on their contribution under smallholder conditions in the tropics. Farm forestry programme in collaboration with other research partners has just begun implementing an on-farm based research project to evaluate the contribution of nitrogen-fixing trees and shrubs to nitrogen nutrition of the trees and associated crops. We are using improved fallow systems as anagro-forestry system, which is increasingly being adopted by the small-scale farmers in the western highlands of Kenya. The objectives of this experiment are to:

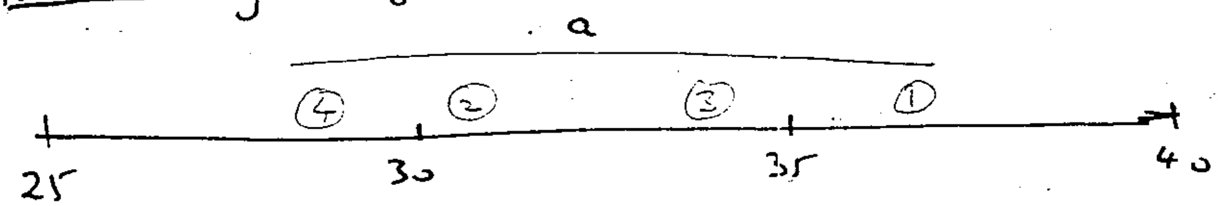
- evaluate *rhizobial* and *mycorhizal* populations in improved fallow systems
- select optimal *rhizobia, mycorhiza* and tree/shrubs for use in improved fallowing
- assess success of establishment and contribution to productivity of trees/shrubs -mycosymbiant systems under improved fallows
- assess success of establishment and contribution to productivity of crops under improved fallows

FREE-HAND SKETCH OF RELATIVE MEAN VARIABLES IN THE METHODS OF INOCULATION

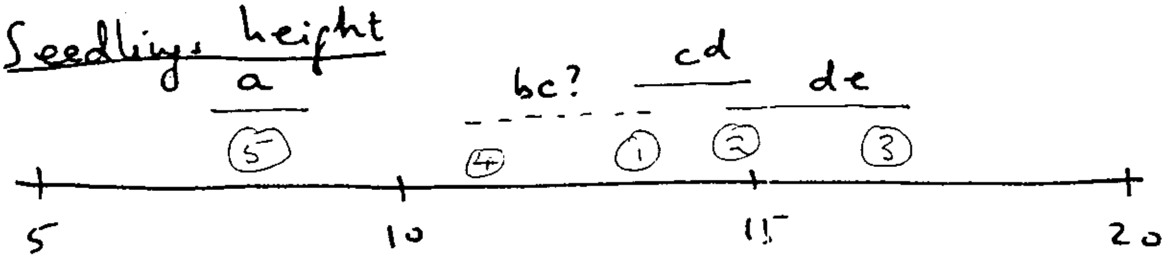
Number of ~~n~~nodules/plant



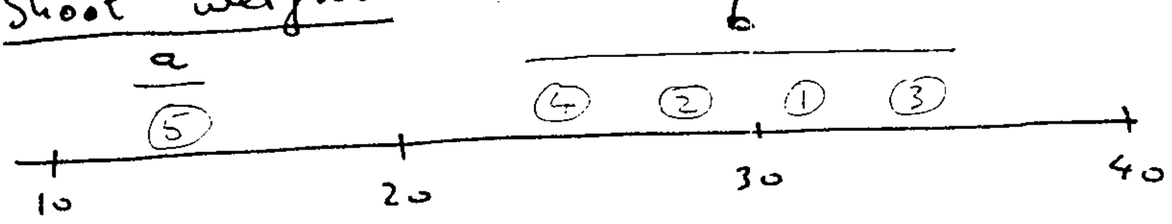
Nodule dry weight



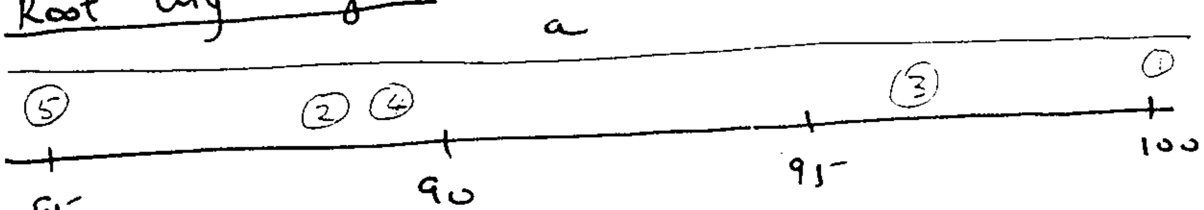
Seedlings height



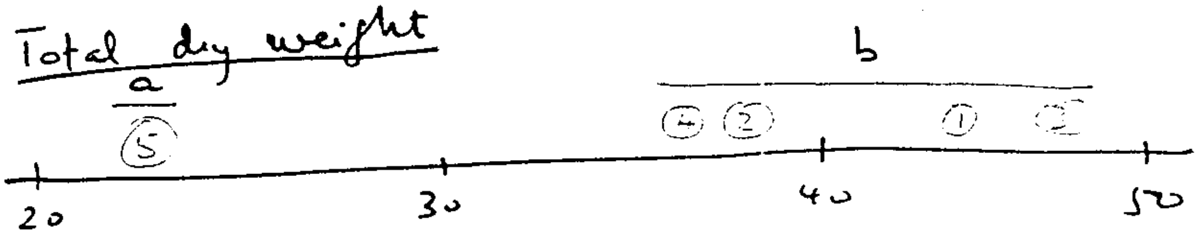
Shoot weight



Root dry weight



Total dry weight



sketch does not have to be exact; use only two (2) significant digits; easy and fast: illuminating for author when used with Table 3 (traditional) and is revealing to reader

Table 11. A row x column representation of the data on DBH of the tree species

Record	Species	Case		DBH	=	Mean	+	C.l	+	E.s	+	P.p	+	Residual
(1)	(2)	(3)		(4)		(5)		(6)		(7)		(8)		(9)
1	1	1		28.00		35.00		-5.00		0.00		0.00		-2.00
2	1	2		30.00		35.00		-5.00		0.00		0.00		0.00
3	1	3		32.00		35.00		-5.00		0.00		0.00		2.00
4	2	1		26.00		35.00		0.00		-8.00		0.00		-1.00
5	2	2		28.00		35.00		0.00		-8.00		0.00		1.00
6	3	1		39.00		35.00		0.00		0.00		7.75		-3.75
7	3	2		41.00		35.00		0.00		0.00		7.75		-1.75
8	3	3		45.00		35.00		0.00		0.00		7.75		2.25
9	3	4		46.00		35.00		0.00		0.00		7.75		3.25

Species labels: 1 – C. lusitanica; 2 – E. saligna; 3 – P. patula

The response (or dependent) variable here is the DBH, usually denoted by y

The treatment (here species) effects are usually denoted by t_i; here $_i$ is for indexing and correspond to the species labels 1,2 and 3

We introduce another index $_j$ to take care of the fact that for each species, a number of observations were made. Thus (1,1) refers to the DBH of 28.00 for *C. lusitanica* i.e. record number 1 whereas (2,2) would refer to record 5: DBH of 28.00 for *E. saligna*.

THE MODEL STATEMENT: that each observation is made up of the mean, treatment effect and residual is stated as follows:

$$\boxed{y_{ij} = m + t_i + e_{ij}}$$

DATA ENTRY: Once we have identified the model, we only need to enter data in columns (2), (3) and (4)

SESSION 8 Wednesday 28 August 2002; 11:50-1:00; Ngugi-Mwangi

FURTHER CONCEPTS IN THE DESIGN AND ANALYSIS OF EXPERIMENTS AND SURVEYS

SUMMARY

In this session we shall:

- *Look at the extension of the CRD to the CRBD;*
- *Introduce the traditional breakdown of Sums of Squares (SS) into Correction Factor, SS due to Treatments, Blocks and Residuals;*
- *Show how a statistical package handles the analysis and*
- *Give a partial breakdown of experimental designs and survey methodologies.*

At the end of the session, the participants should be able to see what to expect in SERIES B.

The complete randomized block design (CRBD)

The example[3] to introduce this design is the mean DBH of trees in plots of 4 provenances of *G. arborera* , 6 years after planting.

Rep	Layout				DBH (cm)				Total	Mean	Effect
I	4	3	2	1	29.4	21.5	30.2	30.8	111.9	27.975	-1.792
II	2	3	1	4	28.4	24.1	38.0	25.0	115.5	28.875	0.892
III	1	4	2	3	35.1	38.0	35.9	20.8	129.8	32.450	2.683

Table 8.1. The layout and mean dbh(cm) of trees of 4 G. arborera provenances

It can be verified that the overall mean is 29.767 and that the block effects are now estimated as shown in the last column. Proceeding as we did for the CRD, we now have the breakdown of the observations as per model for RCBD and the corresponding sum of squares as shown in Table 8.2.

This method of developing the ANOVA although lengthy, nevertheless enables the concepts of effects, residuals and degrees of freedom to be better appreciated. Another disadvantage with it is that one would have to carry enough decimal places to maintain a good level of accuracy. All these shortcomings are reduced by

[3] Adapted from Jayarayam (2000), p86

having short cuts which will be demonstrated in the lecture and are also shown in the Appendix.

Partial Classification of Experimental Designs

- Systematic
- Randomized
    - Experimental material not grouped - Completely randomized design (CRD)
    - Expermental material grouped
        - Complete blocks
            - Single restriction - Randomized complete blocks design (RCBD)
            - Two restrictions -Latin square; Cross-over
            - More than two restrictions -Graeco-Latin square
        - Incomplete blocks
            - Blocks not grouped into replication
                - Balanced and partially balanced incomplete blocks
            - Blocks grouped into replication
                - Lattice
                    - Partially balanced; Balanced
                - Split plot and split block
                - Lattice square
                    - Semibalanced
                    - Incomplete
                    - Balanced
                - Rectangular lattice
                - Youden square
                - Chain block

Some Survey Sampling Techniques

- Simple random sample (SRS)
- Purposive sampling
- Systematic sampling
- Cluster sampling
- Stratified sampling
- Multi-stage sampling

Table 8.2 DECOMPOSITION OF OBSERVATIONS AND SUM OF SQUARES FOR THE RCBD

						Sum of squares			
Observation	29.400	21.500	30.200	30.800		864.36	462.25	912.04	948.64
	28.400	24.100	38.000	25.000		806.56	580.81	1,444.00	625.00
	35.100	38.000	35.900	20.800		1,232.01	1,444.00	1,288.81	432.64
									11,041.12
=									
Mean	29.767	29.767	29.767	29.767		886.05	886.05	886.05	886.05
	29.767	29.767	29.767	29.767		886.05	886.05	886.05	886.05
	29.767	29.767	29.767	29.767		886.05	886.05	886.05	886.05
									10,632.65
+									
Block efect	(1.792)	(1.792)	(1.792)	(1.792)		3.21	3.21	3.21	3.21
	(0.892)	(0.892)	(0.892)	(0.892)		0.80	0.80	0.80	0.80
	2.683	2.683	2.683	2.683		7.20	7.20	7.20	7.20
									44.82
+									
Provenance efect	1.033	(7.633)	1.733	4.867		1.07	58.27	3.00	23.68
	1.733	(7.633)	4.867	1.033		3.00	58.27	23.68	1.07
	4.867	1.033	1.733	(7.633)		23.68	1.07	3.00	58.27
									258.07
+									
Residuals	0.392	1.158	0.492	(2.042)		0.15	1.34	0.24	4.17
	(2.208)	2.858	4.258	(4.908)		4.88	8.17	18.13	24.09
	(2.217)	4.517	1.717	(4.017)		4.91	20.40	2.95	16.13
									105.57

Sessions 9 and 10 Exercises and Computer applications

Wednesday 28 August 2002; 2:15-3:45pm; Break; 4:00-5:00
Resource Persons

Randomisation

A table of random number digits presents a sequence of digits that have been produced through a pseudo-random process. You start reading off the digits from a random starting point. There are various ways of getting random starts such as by use of the sharp end of object (the tip of a pencil, say); you close your eyes and the digit nearest to the tip is the random starting point. The direction you take whether right, left, up, down or diagonally has also to be randomly chosen. You must record the position of the random starting digit as well as the direction or pattern you have chosen. This is so as to enable the supervisor to crosscheck later to ensure that the right procedure has been followed.

The appendix Table A.I is taken from the book by Thomas Little. Suppose you start with the digit 7 in row 2 column 6. Reading to the right you obtain the sequence:

{7, 5, 5, 6, 3, 0, 7, 7, 1, 9, 1, 6, 1, 7, 4, 1, 7, 1, 3, 7, 9, 3, 3, 7, 1, 9, 3, 9, 5, ...}

After reaching the end of the line (where the underlining starts), you could read down one row and then backwards so that you have a zig-zag pattern; you would still generate random digits. Suppose you want to select 5 items, which you number: 1,2,3,4,5. There are several methods you could use.

- Using the above sequence would produce the selection order: 5, 3, 1, 4 and 2.. Note that after selecting 4 items, the fifth one, 2 in this case, is fixed i.e. we say that there are 4 degrees of freedom! Note also the digits that lie outside the {1,2,3,4,5} set i.e. 0 or digits higher than 5 are left out as well as those digits that have appeared before. We say that we are sampling without replacement. There will be cases where we need to carry out swr.

- Instead of choosing from {1,2,3,4,5}, suppose we choose the five items from the set {0, 1,2,3, 4} – we can do this because zero (0) is also a number. This time we select: 3, 0, 1, 4, and then 2 is fixed (coincidence): Adding one (1) to these digits we get 4, 1, 2, 5, 3.

    - *Note* (1) This set is different from that in (a) which was 5, 3, 1, 4, 2; (2) By including 0, you take shorter to complete your selection.

- The method of modulus arithmetic. In division of 8 by 5, we note that 5 goes into 8 one whole time, leaving over 3. Hence, we say 8 mod 5 is 3, which we write as 8 mod 5 =3. When 13 is divied by 5, we see that 5 goes into 13 two whole times, leaving over 3; again here, we write 13 mod 5 = 3. However, when 3 is divided by 5, we note that 5 goes into 3 zero (0) times with remainder 3, and so 3 mod 5 = 3 also. A final example is to see that 5 mod 5 = 0 since 5 goes into 5 one whole time with nothing or zero (0) left over. The modulus of a number is what is left over after you have divided by another number. The following table illustrates some modulo 5:

Number	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Mod 5	3	2	1	0	4	3	2	1	0	4	3	2	1	0

    - Remark that for mod 5, the digits are 0,1,2,3 and 4. We shall see how nicely this relates to the second method of selection.

    - Using this concept then we get from the sequence we started off with the mod sequence would be:... The selection is then 5: 2, 0, 1, 3 then 4 is fixed. When we add one back, we obtain the selection: 3, 1, 2, 4, and 5.

    - *Note*: This time we yet have another random set but we got through a much shorter series.

To recap, with each of the methods, we obtain:

1) 5, 3, 1, 4, 2
2) 4, 1, 2, 5, 3
3) 3, 1, 2, 4, 5

We see that, with the same starting point, different selections are made but that we take shorter time to complete the selection with the latter methods. The price we pay for that is we have to do with a little more arithmetic.

Selection of between 10 and 99 items

What we have seen above is the method of selecting 10 items or less but when we have 11-100 items, there will be need to use *pairs* of digits. Thus if we are to select 13 items using the same starting point, we would read off the pairs:

{75, 56, 30, 77, 19, 16, 17, 41, 71, 37, 93, 37, ...}

You see that numbers in the range 01. 02, ... 13 occur rarely so that to select 13 items this way would take a rather long time. However, using the modulus approach, for the same starting point, would yield.

06, 04, 0,4, 12, 06, 03, 04, 11, 02, 11.06,

From this sequence, we would then select (without replacement), the numbers: 06, 04, 12, 03, 11, 02, ... We would then add one to give us the final selection as:

07, 05, 13, 04, 12, 03, ...

There are other methods used e.g. *Technical Bulletin Series Number 2, November 1993*: **Sampling Cypress Aphids.**

Exercise

1. Find the moduli of the following numbers: 11 mod 5; 11 mod 4; 2 mod 3; 8 mod 4; 33 mod 13.

2. For the example shown above, make a selection of 13 items from the same starting point using the first method. What do you notice about how long you take?

Complete the table for LSD

		C. lusitanica	E. saligna
		30.00 (3)	27.00 (2)
P. patula	42.75 (4)	sed, LSD	
C. lusitanica	30.00 (3)	-	

<u>Exercise</u> Suppose the lengths of shoots for two species have been measured as shown in the following table. You are to complete the xx's:

Case	*E. Saligna*	*P. Patula*	Total
1	4	3	-
2	6	2	-
3	8	-	-
Total	**xx**	**xx**	**xx**
Count	**xx**	**xx**	**xx**
Mean	**xx**	**xx**	**xx**
Species effect	**xx**	**xx**	*xx*

Next complete the following table for decomposing the observations.

Observed		=	Mean		+	Species		+	Residuals	
4.0	3.0		XXX	xxx		XXX	xxx		XXX	xxx
6.0	2.0		XXX	xxx		XXX	xxx		XXX	xxx
8.0	-		XXX	-		XXX	-		XXX	-
xxx			xxx			xxx			xxx	

Then complete the partitioning of the sum of squares.

Observed		=	Mean		+	Species		-	Residuals	
XXX	XXX		XXX	xxx		XXX	xxx		XXX	xxx
XXX	XXX		XXX	xxx		XXX	xxx		XXX	xxx
XXX	-		XXX	-		XXX	-		XXX	-
SS ->xxx			xxx			xxx			xxx	
DF ->xxx			xxx			xxx			xxx	

Finally put this in the format of the ANOVA Table for presentation:

Source	***DF***	***SS***	***MS***	***Fc***	***Ft***	***p-level***
Corrected total SS (CSS)	xx	xxx	-			
Treatments SS, TrSS	xx	xxx	xxx	xxx	xxx	xxx
Error SS (ESS)	xx	xxx	xxx			

Decide whether the mean lengths are significantly different.

Computer exercise

Perform ~~the exercise on the DBH data both on SPSS and~~ GENSTAT and note salient features in the outputs.

TABLE A.1.
Random Numbers

To randomize any set of 10 items or less, begin at a random point on the table anc follow either rows, columns or diagonals in either direction. Write down th numbers in the order they appear, disregarding those that are higher than th number being randomized and those that have appeared before in the series. I you wish to randomize more than 10 numbers, pairs of columns or rows can b combined to form two digit numbers and the same process followed as tha described above.

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

Sessions 11 and 12 Thursday 29 August 2002, J. Chiria 8:40-10:00 and T. Achia 10:20-11:40am

DATA MANAGEMENT

SUMMARY

The two theoretical sessions on Data Management describe the processes involved between problem identification and report writing at the end of the study or investigation. The researcher must be involved in:

- *Planning and designing of data collection format;*
- *Relating data analysis and data collection through the model statement;*
- *Verification and validation of data at field level and*
- *Data preparation, analysis, storage and retrieval*

The practical sessions are planned to take place in between the theoretical sessions so that the relevance of the concepts are demonstrated.

Introduction

The researcher is the one who either singly or as a team member identifies a (social) problem for which a remedial solution is required. The researcher is also the one who will write the report of the study that might give pointers to possible avenues to explore in the future. In between there are several processes that go on which the researcher must not totally abdicate to other persons and for which a certain amount of responsibility should be assumed. Thus from problem identification, to the collection of raw data, to analysis, reporting and archiving and retrievals, must all be properly managed.

Data Management (DM) has of recent gained in importance. Because of the advent of IT we now have e-publishing which demands that the way data were collected and analysed must be part of the paper In the pre-90's this was not a major requirement and in fact it was discouraged because of being bulky but now with the chip technology, you can and ought to reproduce the data collected. In publications such as NATURE, it is required that you must be able to put the data you collected into the public domain. Some donors now require it not only for accountability but also as insurance that what you set to do can be verified. This is important in areas such as molecular biology or protein analysis.

Forestry research takes a long period of time; often running into years. This points to the need for proper documentation of the study right from the beginning to its end. With the use of modern IT tools, the data can now be easily and properly archived and conveniently retrieved.

A '60-40' rule is being proposed here that much more of your time as a researcher should be spent thinking, planning, organizing and supervising the initial stages of the processes of DM. Once this is properly done the latter stages of analysis, interpretation and reporting will follow smoothly.

The Basic Data Format

Raw data that are collected must be prepared before it being fed into the computer, which is merely a machine that processes data to produce information. It is fed through a tape or disk for processing. This is then stored in machine-readable form as data or programs for further use. Alternatively, it can be produced on paper in human-readable form fór people to use.

It is important to realise that the basic data format for (computer) analysis is the *row x column* structure or the familiar worksheet format. In this format, the row takes various names: observation, case, record, whereas the column names are: variable, field, variate

Our problem comes down to this: in designing your questionnaire or preparing your schedules (as in the Exhibits A, B, C), have you thought of about how the data collected will be transformed into this basic *row x column* structure?

The other thing to keep in mind that at or before analysis, data will be sorted, merged and/or split in all ways before they become amenable to analysis. In SPSS this is done through the transformations and in GENSTAT through the **calculate** procedures. This would be an important consideration when dealing with questions where there are multiple responses (illustrate). This fact must be recognized early and taken into consideration at the data format stage.

ICRAF (discussion with Muraya) has developed some simple rules for data preparation: 1. One questionnaire per row, 2. Every bit of data item you require is entered in a column and 3. Where you have multiple response, treat the question as a subtable and introduce a primary key. These three simple rules can alleviate a lot of problems in data format and preparation.

Model Statement

Hand-in-hand with the data format, we also want to have in mind an idea of the skeleton structure or *model* of the underlying factors we want to study. What initial relationships are you looking for? What patterns do you expect to find?

If this can be answered, then the form of analysis becomes evident or at least begins to take shape: we then know that we shall be using a t-test, ANOVA, regression analysis or some other statistical technique. Oftentimes the researcher will collect data and ask the biometrician: "**What statistical test should I use and which statistical package is appropriate?**" We want to avoid this.

So the importance of the basic data format and the model statement lies in the point that there should be reconciliation between data collection convenience with statistical techniques to be used as well as software requirements.

The researcher may be concerned with extracting as much information as possible from the respondent and yet, the analyst may be more concerned with getting the analysis done. So very early in the initial stages of the study, a rapport (round table conference) should be established between the researcher and the biometrician/analyst. Team work is important and can be very illuminating.

Data Preparation

Here we shall not be overly concerned with questionnaire design and the techniques of actual survey. This will be done in SERIES B but here rather, we are concerned with the principles behind these as relates to DM.

During the data collection stage in the field, what has been collected must be *verified* and *validated*. We must check for logical consistency such as:

- **Exhibit A**: Main subsistence crops listed but no farm size indicated
- **Exhibit B**: The name of farmer is *John Kariuki* and sex of head of household is ticked as *F*; was this meant to be for the person responding?
- **Exhibit C**: Questions 2.7, 2.8 and 2.9: do the number of males and females add up to number in household

The data collector or enumerator must scrutinise the source documents thoroughly for obvious errors and blunders. It is only then that the *signed* and *dated* questionnaires can then be passed on to the supervisor or co-odinator at the district, location, province or national level who must in turn check before *signing* and *dating*. Even you as the researcher you must go through the questionnaires albeit on a sub-sample basis.

By merely looking can you spot anything unusual even before you think of data entry or analysis? When you start carrying out the analysis, two plots that are very useful are scatter plots for relationships and line plots for trend (demonstration). For small to moderate sets of data, spreadsheets can be used for data entry while for large set Microsoft Access can come is very handy. But no matter what software package you use, you should not altogether eliminate the need for sensible data scrutiny.

Data preparation should incorporate careful attention to open and close-ended questions. Whether questions are open or close-ended will, among consideration depend on the number of interviews, purpose of survey, time available and objective of the survey undertaking.

Close-ended questions have limited responses that restricts the respondent's replies, which can be single, or multiple responses. They have an implied pre-coding (example Exhibit B: A.1 and A.2).

Open-ended questions are invaluable when you want to find out the general opinion or consensus. But if the respondents' options vary widely, it becomes difficult to capture these and/or summarise them such as when you ask the farmer to list the six most important crops in their order of importance (example: *Exhibit A question B*). The open-ended questions can give you a lot of important feedback and details about the respondents' attitudes, perceptions and opinion.

With one type of open-ended question, you have a pre-determined list of answers that you expect to receive. You assign a code for each name or response. You might have a list of crops with you, which you already have pre-coded. There is now some effort required on the part of the enumerator to always refer to the list, identify the code and write this down in the '**Leave blank**' column for the data entrant to enter. Besides this taking long to record, what other things could also go wrong in the process? [Imagine 10 open-ended questions administered to 100 respondents. If you must summarise 100 responses for each of the 10 questions you must have a lot of time available at hand].

The other type of open-ended question has a multiple or combination of answers. A question such as problems associated with the establishment of xxx... "**Exhibit A.** J. ADOPTION BY FARMERS" will generate responses that are not so quickly classified. In order to better analyse the results, you could develop a list of categories into which any one *problem* could fall. You should limit the number of categories.

Generally, as much as possible, questions should be pre-coded. This demands a lot of effort on the part of the chief investigator or team leader. But once this is sufficiently done, subsequent data collection and more importantly, data analysis, becomes fairly straightforward. The open-ended questions should be placed at the end to give more flexibility and more room for verbatim responses.

Each questionnaire must be uniquely identified through a sheet or questionnaire ID. This will help track down problems during data cleaning as well as to flag cases of particular interest during analysis.

Nichols (1991) advises thus: " A well designed form is a vital research tool. Use 'closed' questions with a limited range of answers, where the answers may be varied, for finding out about attitudes and experiences. In general, closed for questions are appropriate for large structured surveys and open questions for preliminary research". There will be need to strike a proper balance.

Production of tabulation plan or dummy tables

You should also generate dummy tables or tabulation plan. This will help the data analyst to quickly identify the variables from your questionnaire that need to be cross-tabulated. For example from **Exhibit C** if it is important to know whether type of crop grown is related to farm size, this should be indicated by the researcher *at the beginning of the study.* The analyst will then know that the variable, *farm size* must be first classified into e.g. 1 – small, for less than 2 acres, 2 – medium, for between 2 and 5 acres and 3 – large, for 5 or more acres. Then the frequency of each crop can be tabulated for each category of farm size.

The investigator should be able to indicate through a sketch of the table or by title, what table is required to appear in the paper or technical report together with where the information is to be picked up from on the questionnaire. If this cannot be done then this could be an indication that you do not need to waste time collecting that piece of information.

Data Analysis

Analysis can be interpreted as looking at data in all possible ways. Analysis delivers the value from survey data. Whenever you have a new dataset, further data verification and cleaning must be done on the computer. Just like what was done in the field, you look at large and small values as well as the 'average' value. You then run a series of cross-tabulations before doing further analysis. The quick cross-tabulations can help to identify:

- Inconsistent relationships (female whose relationship to head of HH is 'son')
- Unexpected averages and outliers and
- A large number of missing values.

You must keep track of the analyses you run. When performing complex analyses, you must keep a record of the procedures you perform or the way you created new variables. This will help you re-conduct your analyses if any questions arise. It will also help you write your report. Keeping a record for example of the SYNTAX in SPSS will enable you to run fresh analyses more fully just by making a few changes only (This will be demonstrated for SPSS).

Remember that as we have seen on the topic on averages and spread of data, different statistical procedures are appropriate for variables depending on what you want to achieve and the level (or scale) of measurement of the variable.

Data Storage and Retrieval

We have seen that the output of computer processing can be in *machine-readab*le form for further use. It can also be in *human-readable* form for needs of people. Output can further be classified as hard-copy implying permanency or retainable over an indefinite period of time as printed reports, plotted diagrams, to tape or disk. It can also come out as soft-copy, which lasts for a short period of time only; an example is information displayed on the monitor.

In the end, the data that have been analysed and reported on, must be archived and liberated or disseminated. The storage, retrieval and liberation into the public domain must be supervised or managed properly by designated person or institution. This person /institution needs to specify where the raw-data have been achieved and how they can be documented and made publicly accessible. The raw data needs to be stored so that in case of verification or a re-analysis it can be retrieved. But of course, the value of all this depends on: How 'good' were the data collected in the first place?

Exhibit A

A socio-economic survey on the development of *Moringa oleifera* and *M. stenopetala* tree to provide valuable products

B. ON-FARM ENTERPRISES

Farm size ……………………………………………………….(acres)

Main on-farm sources of income.

1……………………………………… 2……………………………

3……………………………………… 4……………………………

5……………………………………… 6……………………………

Main subsistence crops grown (in order of importance)

1……………………………………… 2……………………………

3……………………………………… 4……………………………

5……………………………………… 6……………………………

E. MANAGEMENT PRACTICES OF MORINGA SPECIES

After planting, what management regimes do you supply?

REGIME	REASON	FREQUENCY
Water		
Weeding		
Pruning		
Spraying		
Treatment		
Other		

J. ADOPTION BY FARMERS

Are there problems associated with *Moringa* establishment?

ACTIVITY	PROBLEMS
Establishment	
Seedlings	
Purchase of seed	
Sales of products	
Others	

Exhibit B

Assessment of land degradation and the impact of the existing pract ces on natural resource components

Farmer's name ____________________ Name of respondent ______________
Relationship of respondent to farmer ____________________

A. GENERAL INFORMATION

A.1 Sex of head of household F □M □

A.2 Age (years) <21 21-3□ 3□0 4□0 >□ □

A.4 Marital status: □ single □ married di□edwidow □ □

If married, how many wives do you have?______________
If married woman, is your husband married to other wives?______

A.5 Occupation:
Farmer □ Pastoralist □ Farmer and pastoralist □
Trader □ Employee □ Others (specify)

A.9 What role does each member of the family play in natural resource management?

Children ____
Teenagers (female)______________________________________
Teenagers (male)_______________________________________
Adults (women)____
Adults (men)____

C. LAND USE

C.2 What do you understand by the terms:

Land degradation ____
____
Desertification____
____
Deforestation____
____
Soil erosion_____ _______________________________________
____
Sustainable development__________________________________
____

Exhibit C

APPENDIX 1: QUESTIONNAIRE FOR GENERAL SURVEY ON ... PRODUCTION AND UTLISATION IN ... DISTRICT

2.0 SECTION B: PERSONAL DATA & GENERAL INFORMATION

2.7 Number of members of your household……………………….

2.8 Number of male members………………………………………

2.9 Number of female members……………………………………

2.14 What is the source of income to your family?………………...

4.0 SECTION E: LIVESTOCK OWNED

4.1 What livestock do you own and in what number?

Livestock	Number
Cattle	______
Sheep	______
Goats	______
Donkeys	______
Camels	______
Others (specify)	______

4.2 What are your major problems regarding livestock husbandry (List in order of importance)……………………………………………………………

4.3 What are the solutions to these problems?…………………………….

5.0 SECTION D: AVAILABLE NATURAL RESOURCES & THEIR USAGE

5.4 What crops do you grow?………………………………………….

Exhibit C (continued)

Selected Records from: "Appendix 2. Raw Questionnaire Data from ….Division, ….District"

NAME	NO. MALES	NO. FEMALES	SOURCE of INCOME	CATTLE	FARM SIZE (ACRES)	LIVESTOCK PROBLEMS	TYPE OF CROPS GROWN
Aaaa	1	2	Farming	0	0	Fodder, Water, ECF, Money	Millet, Sorgum, Green grams, Groundnuts
Bbbb	3	2	Livestock, Honey	2	None	Diseases, Water, Lack of food	Millet, Sorgum, Groundnuts
Cccc	4	3	Charcoal, Farming	0	None	Fodder, ECF	Millet, Sorgum, Groundnuts
Dddd	7	4	Livestock	15	5	Drought, Cattle rustlers	Millet, Sorgum, Pawpaws, Maize, Beans, Groundnuts
Eeee	2	4	Employment, Goats	2	None	Diseases, Drought	Millet, Sorgum, Beans, Green grams, Groundnuts

Sessions 13 and 14: Computer applications

Wednesday 28 August 2002, *Resource Persons*

Participants to be broken into groups for discussion of the **Exhibits.** Each group leader will choose someone to summarise the groups' discussion on the Exhibit.

A small sample of actual survey data should be entered on the computer and analysed by each group to show understanding of the major concepts of DM introduced in the workshop.

Session 15: Wrap-up session: Review and Appraisal

REFERENCES

Box, G. E .P., Hunter, W.G. and Hunter. J. S. (1978). *Statistics for Experimenters: An introduction to Design, Data Analysis and Model Building*, John Wiley and Sons Inc.

Coe, R. (1998) Participatory on-farm experimentation in Agroforestry: Experiences and the role of biometrics. Invited papers presented at the XIXth International Biometrics Conference, Cape Town, South Africa, 14/12- 18/12.

Coe, R and S. Franzel. (1995). Design of on-farm agroforestry experiments. ICRAF Traiing Notes.

Cochran, W.G. and G.M. Cox (1964). *Experimental Design*. John Wiley and Sons, Inc. New York.

Day, R. *et. al.* (1993). Sampling Cypress Aphids, Technical Bulletin Series, Number 2, November, IIBC.

Jayaraman, K. (2000). FORSPA–FAO PUBLICATION: "A Statistical Manual for forestry Research.

Kaudia, A. A. (1996). The Diffusion of Social Forestry in Semi-Arid areas: A Case Study in Kitui District, Kenya. Thesis submitted to the University of East Anglia for the Degree of Doctor of Philosophy, U. K.

LeClerg, E. L. W. H. Leonard, and A. G Clark (1962), *Field plot technique*, Burgen Publishing Co, Minneapolis, Minnesota

Little, T. M. and Hills, F. J. (1978). *Agricultural Experimentation: Design and Analysis.* John Wiley and Sons, Inc.

Mead, R. (1988). *The Design of Experiments*. Cambridge. Cup.

Muraya, P. (2002). Data Management training course, ICRAF

Nicholas, P. (1995). Social Survey Methods. A field guide for development workers. Development Guidelines No. 6.

William, E. R. and Matheson A. C. (1994). Experimental design and analysis for use in tree improvement.