STATISTICAL AND GRAPHICS SOFTWARE FOR DATA ANALYSIS

K.Muralidharan

INTRODUCTION

Statistics is a rapidly growing subject which came into existence as a branch of science only in the twentieth century. The availability of cheap computing facilities and software have revolutionized the application of statistics in every aspects of modern living. Statistical planning and evaluation has been routinely used by industries to make advances in processing and statistical quality control of products makes them reliable.

There are several definitions for statistics - collation of data in primitive sense to, as a branch of science (pure and applied) of creating, developing and applying techniques such that the uncertainty of inducting inferences may be evaluated. Generally, statistical application involves formation of a hypothesis, collection/ generation of data and finally, based on an appropriate test a decision to reject or not to reject the hypothesis.

The hypothesis may be concerned with comparison of groups (treatments), the value of a population, association/relation between two or many characters (variables)/ individuals etc. The relevant data for testing the hypothesis may either be collected from a 'sample' of observations or generated from an 'experimental design' as what the case may be. To test hypothesis related with time, often secondary data is used.

A number of statistical procedures are used in the field of harvest and post harvest technology. However, some of the most useful techniques are seem to be overlooked. The purpose of this note is to generate a discussion on the appropriateness of certain advanced statistical techniques as a research aid of harvest and post harvest technology of plantation crops.

DATA COLLECTION

As mentioned earlier, data is obtained mainly from a sample or from an experimental design. A 'sample' is a part of a population; all possible values of the variable comprise the population. The 'experimental design' deals with planned inquiry to obtain new facts, or to confirm or deny the results of previous experiments. Broadly, the purpose of an experimental design is either for comparison of treatments or for describing a response function whereas in a sample survey one is interested in estimation of population values (parameters).

As in experimental designs, a number aspects are to be considered while choosing a 'sample design'. Ways in which the sample units is drawn termed as 'sample designs'. Defining the sampling units and population is fundamental to any sample survey. This process should enable one to locate and identify the units in the population if required. Based on the nature of sampling units, the sample design is decided. The most simple and commonly used sample designs are (1) simple random sampling, (2) stratified sampling, and (3) multi stage (cluster) sampling. Systematic sampling is also popular. Cochran (1984) may be referred for

a detailed account on these and related topics. Sampling is also employed in experimental design to collect data from a fraction of a plot.

SOFTWARE FOR DATA ANALYSIS

Many statistical application software are in use. Some of them are very simple and users friendly as well. It is not possible to make a comparison among all the commercially available software. Neither the author is experienced enough to comment on the same. Interested readers may consult the product reviews published by computer magazines. The Indian Agricultural Statistical Research Institute, Library Avenue, New Delhi-12 had released two statistical software - one for data analysis (SPAR1) and another for selection of block designs (SPBD). At CPCRI, we use SPSS for data analysis and AXUM for graphics.

STATISTICAL PROCEDURES

Preliminaries:

The initial steps in data analysis include, summarization of data, checking for unusual values (outliers or wrongly entered values) and identifying the pattern if any. The SPSS procedure EXAMINE offers a variety of ways for a detailed examination of the data. Before the actual trial of a hypothesis, there is much evidence to be gathered and sifted. If necessary the hypothesis and/or the methods of testing may have to be changed.

With regard to a quantitative data, we may first obtain the frequency distribution. Most of the frequency distributions are characterized by its mean and variance. Since the data we come across is often a sample from an infinite or a finite population, there exists an amount of uncertainty on our inference. To measure the uncertainty we utilize the concept of probability. A simple definition of probability is the relative frequency. The frequency distribution of values of a variable in a 'population' is called the probability distribution of that variable. A large number of quantitative variables are found to follow the normal distribution; a great many may follow after appropriate transformation. Important discrete distributions are binomial, multinomial, and Poisson.

The values of an estimator will vary from sample to sample. On knowing the distribution of the estimator, the upper and lower bounds within which the values are expected to fall with a preassigned confidence (say 95%) can also be constructed - the confidence interval. The frequency distribution of values of an estimator (statistic) is referred as the 'sampling distribution' (e.g., F, t, Chi-square). We utilize these distributions to express our confidence on the estimates obtained from a sample. We also utilize the fact that the sum of the observations of a sample often follows a normal distribution irrespective of the parent distribution (the central limit theorem).

The estimated values of a parameter may have to be tested for its agreement with the value proposed by the experimenter (the null hypothesis). The null hypothesis is tested against an alternative hypothesis. There are two types of errors in this process: The Type I error - rejecting the null hypothesis when it is true and the Type II error - accepting the null hypothesis when it is false. The probability of Type I error is called the level of significance (generally fixed as 5% or 1%). The complement probability of the Type II error (the probability of a correct decision) is termed as the power of the test. The most

frequently used tests are the t-test, paired t-test, F-test and Chi-square test.

Regression analysis:

An important use of statistics is in the exploration of relationship between the response variable(s) and causative variables: The linear relationships in particular. A measure of (linear) association between two variables is the coefficient of correlation. The explanatory model relating the dependent variable and the predictor variable is termed as the regression model. Regression analysis is one of the most versatile data analysis procedures. The coefficients of a linear regression model is estimated by using the least square method. It is assumed in a regression analysis that the observations are independent and normally distributed with equal variance. By looking on residuals the validity of these assumptions may be assertained. SPSS offers a number of options for the 'residual analysis'. These include different scatter plots, the Durbin-Watson statistic etc. If the assumptions of linearity and homogeneity of variance are met, there should be no relationship between predicted and residual values. If the value of Durbin-Watson statistic is not close to 2, one may suspect that the observations are not independent. When building a regression model, it is important to identify points that are influential. The important influence measures is the 'leverage'. Another aspect to be looked into is the possible high correlation among few predictor variables.

Linear equations are the simplest to fit an observed relationship. But in many situation, it may be necessary to fit nonlinear models. There are different algorithms for the estimation of nonlinear models and explained in Draper and Smith (1981).

When the dependent variable can have only discrete values (say, present or absent), the 'logistic regression model' will be used. In this case, the parameters are estimated using the 'maximum likelihood method'. That is the estimates that make our observed results most 'likely' are selected.

Loglinear models:

Analysis of contingency tables is one of the most debated topics in statistics. When more variables are included in a cross classification table, the number of cells rapidly increases and make it difficult to analyze the association among the variables by examining only the cell frequencies. The loglinear models are formulated for this kind of situation. Log linear models are similar to multiple regression models - the classification variables are the independent variables and cell frequency is the dependent variable. Interested reader may consult McCullagh and Nelder (1989) for more details.

Analysis of repeated measurements:

In many experimental situation, measurements of the same variable are taken at several occasions (time) for each subject (experimental units). The observations are then cannot be considered as independent. A number of procedures of analysis are proposed to deal with such situations of which the univariate (split-block) analysis of variance and repeated measurements multivariate analysis of variance are easy to use. In split-block ANOVA, the error is partitioned as variance among experimental units with regard to averages (or sums) over time and variance among observations within experimental units. This ANOVA is different from that of a split-plot for one additional source of variation viz., replication X time.

Since the 'sub-plot classification' (time) is not randomized, it become necessary to test the assumption of equal variances (at each time) and equal covariances (between pairs of time) to justify the use of split-block analysis. The chi-square test due to Box may be used to test this assumption of symmetry. The Repeated measurements MANOVA is identical with the usual multivariate analysis of variance (MANOVA) and the treatment x time interaction effects are tested based on an appropriate 'sets of contrasts'. The SPSS procedure for MANOVA for repeated measurements will perform both these analyses. More details on analysis of repeated measures may be obtained from Crowder and Hand (1990).

Survivorship analysis:

The statistical analysis of life time of products, storage time etc. are often come across in Engineering and Food Science. Mathematically, the life time or survival time is a non-negative valued variable. The probability of an individual survive till a specified time is termed as survival function or reliability function. The instantaneous rate of death or failure at any time is defined as the hazard function. A problem one could encountered in survivorship studies is that the experiment might have completed before determining the life time of some of the individuals. Data of this kind are called 'censored data'. Special statistical techniques are needed for the analysis of censored data. Kaplan-Meier estimation of survival rates and the Mantel-Haenszel (logrank) test for the comparison of estimated survival curves (for different levels of the predictor variable) are the most widely used devices in survivorship analysis (Haris and Albert, 1991). When the predictor variable is continuous, special type of regression model, 'Cox's proportional hazards model' is used. The parametric approach to deal with censored data is to employ parametric families of lifetime distributions and extends models such as the exponential, Weibull, and lognormal models to include regressor variables (Lawless, 1982).

Nonparametric methods:

When the assumptions (on the distribution from where the sample is drawn or regarding the population parameters of two or more population) of test is expected to be violated, one may use nonparametric tests; a majority of non-parametric tests are based on ranks of the sample observations.

REFERENCES

Cochran WG (1977). Sampling Techniques (3rd Edn.). Wiley, New York.

Draper N and Smith H (1981). Applied Regression Analysis (2nd Edn.). Wiley, New York.

Haris EK and Albert A (1991). Survivorship Analysis for Clinical Studies. Marcel Dekker Inc., New York.

Lawless JF (1982). Statistical Models and Methods for Lifetime Data. Wiley, New York

McCullagh P and Nelder JA (1989). Generalized Linear Models. Chapman & Hall, London.