PRELIMINARY TESTING IN THE ANALYSIS OF DATA

by

SAVITRI ABEYASEKERA

(Statistical Unit, University of Colombo)

Thesis submitted for the Degree of Doctor of Philosophy

University of Colombo, Sri Lanka

August 1981

SUMMARY

A common requirement when data are collected for purposes of research is to make a preliminary determination of whether the data satisfy assumptions to be made in the subsequent analysis. This thesis is a study of the application of certain preliminary tests in relation to the effect they have on the particular inferential procedure under consideration. The situations investigated are the following.

(1) Location estimation when sampling from a possibly contaminated population.

Investigated here is the specific situation when data emanate from a normal distribution but may include a single observation (an outlier) from another normal population shifted in mean but having the same variance. Rejecting the most extreme observation from the sample following a test of discordancy is considered. When the direction in which contamination may occur is known, the population mean is estimated with smallest mean squared error by always rejecting the most extreme observation, while in the case when the direction of contamination is unknown, certain threhold values to be used in the test of discordancy are recommended.

(2) Parameter estimation in a simple linear regression model $E(y)=\alpha+\beta x$ in the possible presence of an outlier.

The study here is similar to that above. From amongst observations made on the dependent variable, the one yielding the largest absolute studentized residual is subjected to a test of discordancy and the resulting effect on properties of the parameter estimators

investigated for varying levels of contamination. It is found that the estimation of β is most affected when the contaminating observation occurs towards the extremes of the x-range and in such cases certain thresholds are recommended for use in the test of discordancy. Several configurations of values specified for the independent variable are also investigated but this is found to be of little consequence.

(3) Performance of the sequential probability ratio test (SPRT) in the possible presence of an outlier.

Testing two simple hypotheses concerning the mean of a normally distributed population, contaminated by a single observation from another normal population shifted in mean is considered. The application of a preliminary test of discordancy at various stages of the sequential procedure and its resulting effect on the probabilities of type I and type II error, as well as the average sample number, is investigated. It is found that the best procedure to adopt is first to sample at least 10 observations, reject the most extreme one and then carry out the SPRT on the remaining observations. Subsequent observations, if taken, also need to be tested individually for discordance as they enter the sample using significance levels that are neither too high nor too low.

(4) The desirability of adjusting for residual effects in a cross-over design.

The effect on properties of treatment effect estimators following a test of significance on the residual effects is studied. Analytical results for the bias and mean squared error of the resulting estimators are derived, and the corresponding numerical results, studied over varying degrees of the strength of residual effects, showed that it

was always best to adjust for residual effects in order to make the estimation of treatment effects most precise.

(5) The desirability of covariance adjustments.

A study is made of certain criteria that may be used in determining whether or not to make a covariance adjustment when estimating treatment effects in different design models. The performance of the criteria is measured in terms of the resulting effect they have on properties of the treatment effect estimators. Here again the procedure of always making a covariance adjustment is found to be the most appropriate action to take when supplementary information on a concomitant variable is available.

(6) Efficient estimation of the parameters of the von Mises distribution.

Measures of location and scale of circular data drawn from the von Mises distribution are studied. Properties of the maximum likelihood estimators of the parameters of this distribution are compared with simpler estimators analogous to those used on the real line. The latter are found to be preferable in most cases.

The investigations described above were made mainly through the use of Monte Carlo methods. The performance of estimators in the different situations were measured in terms of their bias and mean squared error values since reducing these would lead to the most precise estimators. An interesting feature that emerged was that very often the application of a preliminary test was not required and greatest precision in the estimation procedures was obtained by adopting an "always take

action" approach. In the case of the studies on outliers, the action was to reject the most extreme observation while in the two design studies, the action was to make the appropriate adjustment. This is fortunate as the experimenter is not burdened with complicated data screening procedures (at least for the situations described above) before undertaking his major statistical analysis.