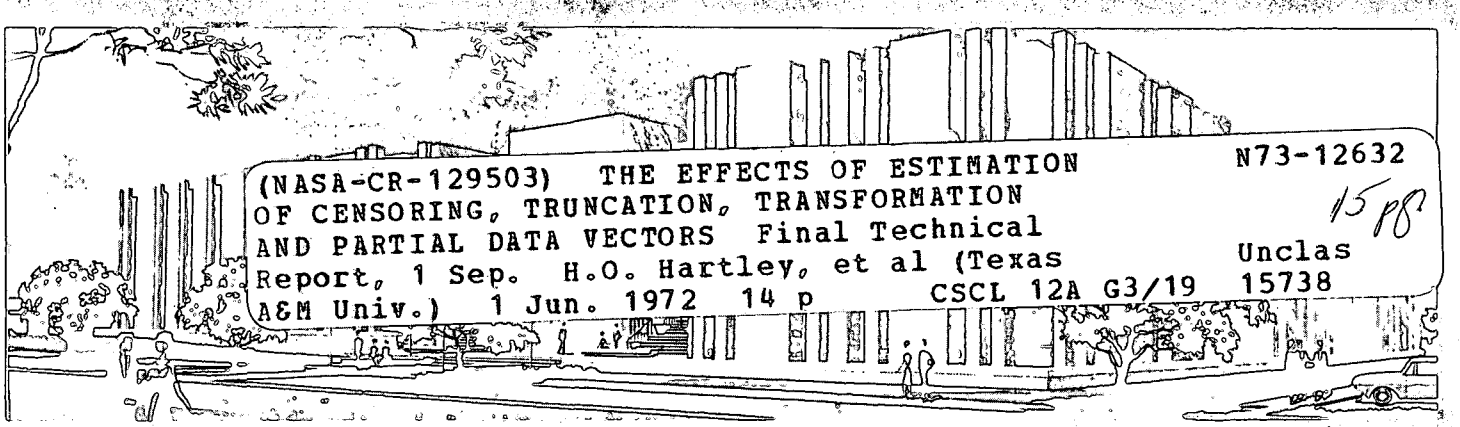
TRA RCH

TRA RCH

TRAINING & RESEARCH

GRADUATE
INSTITUTE
OF
STATISTICS



TEXAS A&M UNIVERSITY • COLLEGE STATION

FINAL TECHNICAL REPORT

September 1, 1969 to June 1, 1972

The Effects on Estimation of Censoring, Truncation, Transformation and Partial Data Vectors

National Aeronautics and Space Administration

Manned Spacecraft Center

Grant # NGR 44-001-095

H. O. Hartley and W. B. Smith
Institute of Statistics
Texas A&M University
College Station, Texas 77843

FINAL TECHNICAL REPORT

The Effects on Estimation of Censoring, Truncation, Transformation and Partial Data Vectors

I. Introduction

The purpose of this research was to attack statistical problems concerning the estimation of distributions for purposes of predicting and measuring assembly performance as it appears in biological and physical situations. Various statistical procedures were proposed to attack problems of this sort - that is to produce the statistical distributions of the outcomes of biological and physical situations which, in general, employ characteristics measured on constituent parts. The techniques used in this grant are described in section II wherein each of the technical reports are briefly abstracted.

During the indicated performance period nine technical reports were issued and many formal presentations were made at meetings of learned societies. Most of the technical reports have been published in scholarly journals and give indications of the several methods used in the solution of the various problems described herein.

/

II. Technical Reports

1. "Optimum Incomplete Multinormal Samples", by R. R. Hocking and W. B. Smith (Technometrics 14, 299-309, May 1972).

When sampling from a multivariate normal population the researcher is frequently forced to cope with incomplete data. The data may be incomplete because not all elements of the multivariate observation vectors are recorded on every occasion (the missing data problem) or because in some cases certain linear combinations of the elements of the observation vector are recorded rather than the individual elements. Such data may arise due to peculiarities in the data collecting procedure or there may be economical or physical reasons for collecting the data in incomplete form. For example it may be costly to make certain measurements on some experimental units, so that these measurements are not taken each time, whereas less costly measurements are made.

A maximum likelihood solution was described recently for the problem of estimating the parameters of a multivariate normal distribution with incomplete data (Hocking and Smith (1968)). In addition to developing the estimates, expressions were obtained for asymptotic variances of these estimates. These expressions made it possible to assess the effect of incomplete data on the estimates. This technical report considers the problem of designing the data selection procedure to intentionally yield incomplete data but at the same time give desired

precision while satisfying certain other requirements. Specifically we consider the problem of minimizing the cost of gathering the data subject to the requirement that the parameters or functions of the parameters are estimated with desired precision.

2. "Lognormal Parameter Estimation from Censored Data", by W. B. Smith, C. D. Zeis and G. W. Syler (Presented at the National Meetings of the American Statistical Association, 1969).

This research report deals with the aspects of maximum likelihood estimation of the three parameters of a lognormal distribution from censored (Type I) data. The existing techniques are discussed: Harter and Moore (1966), an iterative scheme for solving the simultaneous likelihood equations; Tiku (1968), a direct technique for solving the modified likelihood equations. A further development due to Syler (1968) with Tiku's procedure was also considered, and a new type of estimation procedure based on both convex and non-convex programming algorithms is outlined. A numerical scheme for determining the elements of the Fisher information matrix in each of the above situations, and hence the covariance matrix, is presented. The small sample bias of these estimates are determined. A Monte Carlo simulation is carried out in which the existing techniques are compared on the basis of expected biases, and variance-covariance matrices are given

for specific parametric situations. In addition a simulation is carried out in which random samples are generated from the three parameter lognormal distribution, and these parameters are estimated using the Hartley-Hocking convex programming algorithm. It should be noted that the likelihood function being considered is not convex, but the procedure is carried out in an attempt to achieve a reasonable starting value while the other iterative procedures, to provide a "measure" of the non-convex objective function, and to provide a contrast to the non-convex programming algorithm solution (Hartley, George and LaMotte (1969)) which when implemented on the computer should give a reasonable estimation procedure.

Briefly this non-convex programming procedure depends on the separability of the objective function into two additive functions one of which is convex the other which is non-convex. In this method, the Hartley-Hocking convex programming algorithm is applied to optimizing the convexity of the objective function for a constant value of the non-convex part. This is done for each iterate of the non-convex part thus providing an "optimal" value of the convex procedure at each stage of the non-convex procedure. Thus if an iterative procedure can be developed so that the non-convex portion approaches some optimal point, then the optimality is guaranteed for the entire objective function. This procedure currently finds only a relative optimum point, that is, not necessarily a global optimum. Implementation of

this procedure is not yet complete, however the results on the existing techniques rather than those of mathematical optimization, are reported on in this technical report.

3. "A Computer Program for the Analysis of Variance Models Based on Maximum Likelihood", by H. O. Hartley and W. K. Vaughn (This report appeared as a chapter in the Snedecor Volume published by the Iowa State Press, 1971).

In this report a computer program is presented for an analysis of variance of unbalanced data assumed to arise from a "mixed model". The analysis is based on the principle of maximum likelihood estimation developed by Hartley and Rao (1967).

A general mixed analysis of variance model involving both fixed and random factors with interaction can be given by

$$Y = X\alpha + U_1 b_1 + \ldots + U_c b_c + e \,, \qquad (1)$$

where X and U are matrices of known fixed numbers, α is a vector of unknown constants, b_1 is a vector of independent variables from $N(0, \sigma_i^2)$, and e is a vector of independent variables from $N(0, \sigma^2)$. Hartley and Rao give solutions to the likelihood equations resulting from repeated sampling on this model. These are the familiar least squares results

$$\tilde{\alpha} = (X'H^{-1}X)^{-1} (X'H^{-1}y)$$

$$n\tilde{\sigma}^2 = y'H^{-1}y - (X'H^{-1}y)' (X'H^{-1}X) (X'H^{-1}y) , \qquad (2)$$

where H is the variance-covariance matrix of y and

$$\gamma_i = \sigma_i^2/\sigma^2 .$$

Using the method of steepest descent one would then be able to solve the likelihood equations for γ_i as well. A full explanation of the formulation is given in the paper by Hartley and Rao, however implementation of the procedure has just now been accomplished and is the subject of the report which will appear. Moreover, the procedure will be used and extended to the calculation of the variance-covariance matrix of the component of variance estimates. This extension will be carried out using statistical differentials (truncated Taylor series expansions).

4. "Statistical Simulation Procedures", by R. N. Tremelling, W. B. Smith, L. J. Ringer and J. L. Oglesby (Submitted for publication).

Ringer and Suharto (1968) extended a standard Monte Carlo simulation technique by an employing stratified random sampling procedure. A stratified Monte Carlo simulation is based fundamentally on stratified

sampling; that is, if you are observing a function of many input variates, a standard procedure simulation would be to generate the entire multivariate vector, calculate the function, record its value and compare it to design standards, and then repeat the procedure. An empirical distribution would be generated thereby. Stratified Monte Carlo simulation is a procedure which partitions the ranges of the input variates into equal probability intervals. Then, instead of sampling from the entire range, each combination of sub-ranges is sampled. Improved (reduced variance) estimates of the cumulative distribution function are obtained. This report includes an illustration from an electronic design with a sixteen input variates, as well as, an extension from the univariate case to the multivariate situation.

5. "Confidence Regions for Variance Ratios in Variance Components Model", by A. S. Al-Barhawe and H. O. Hartley (March 1971).

The situation frequently to be met in applied statistics is as follows: We have a set of data arranged in a particular type of classification and described by the linear function of effects on various classes and sub-classes. Generally this model is that which Eisenhart called Model II in which all elements except the mean are regarded as random variables, although it may be frequently called the mixed model in which certain of the effects are regarded as fixed rather than random variables.

The unknown constants (i.e., the variances within compartments) are called components of variance and point estimates of these components are now used in many fields of research. This report includes confidence intervals for the ratio of the variances of interest, as well as, the point estimates. Comparisons are made between this technique and that of Wald, and specific distributional properties of these variances are spelled out for several cases.

6. "Incomplete Data Analysis", by H. O. Hartley and R. R. Hocking (Biometrics 27, 783-825, December 1971).

It is a common-place experience of many practicing statisticians to be frequently faced with the task of "analyzing incomplete data". The evil of incompleteness is most frequently encountered with data routinely collected and while the occurrence of incomplete data may be regarded as a necessary evil to those associated with data collection, the literature provided ample evidence of considerable efforts to provide solutions to such problems. This paper provides what is regarded as a simple taxonomy for the occurrence of incomplete data to compose a framework of coordinating the bewildering multitude of species of incompleteness. Moreover, two unified methods of analysis are proposed for certain aspects of this taxonomy. These two techniques represent first a new application into what is called the "group" data situation,

and secondly to a unified method of finding maximum likelihood estimates based on the incomplete data observations (originally espoused by Hocking and Smith (1968)).

7. "Maximum Likelihood Analysis of Balanced Incomplete Block Models", by M. H. Kutner (July 1971).

Several methods exist for obtaining point estimates of variance components with unbalanced data. For example, Henderson gives three unbiased estimation techniques, Hartley and Rao describe a procedure for obtaining maximum likelihood estimates of fixed effects and variance components in a mixed analysis of variance model, and LaMotte considers near maximum likelihood procedures.

This paper examines closely the likelihood equations resulting from analyzing balanced incomplete block designs. Structural forms of the likelihood equations are developed, resulting in increased computational efficiencies and ease of investigation of small sample properties. For completeness, a matrix analysis of the fixed effects model is given. The relationship between maximum likelihood and analysis of variance is investigated in the mixed model, and the completely random model is also studied. Numerical comparisons are made with other procedures.

8. "Wishart Variate Generator", by W. B. Smith and R. R. Hocking (Accepted for publication in Journal of the Royal Statistical Society, Series C, 1972).

This paper consists of a computer algorithm for generating pseudo-random matrices, whose distribution is Wishart. That is, random sample covariance matrices from vectors distributed as multivariate normal are produced. The procedure used is efficient and follows the Bartlett decomposition technique.

9. "Sequential Control Chart Methodology", by J. M. Lucas (Accepted for publication in Technometrics).

A modification of the "V" mask sequential control chart is proposed. In this modified scheme, a parabolic section is included in the mask to provide better performance when a process undergoes a large change in the mean from goal conditions. It is shown that the modified "V" mask can be implemented either by conventional graphical form, or in an algorithm form suitable for a digital computer. Average run lengths are given for a typical range of circumstances. It is also shown that the conventional Shewhart chart is better than a sequential chart for the specific purpose of promptly detecting very large shifts of the mean from goal conditions.

III. Summary and Conclusions

As indicated by the abstracts given in section II several types of statistical estimation and optimization problems were considered. These problems range from Monte Carlo simulation of statistical distributions resulting from certain physical phenomena, to the estimation of statistical components of variance which appear when data is gathered under a specific structural form, to missing data problems both in estimation and in optimally designing experiments to yield missing data while giving desired precision on the estimates. Thus the research funded by this grant covers several major areas in applied statistics. It should be pointed out that each of these areas finds immediate application to real world problems.

For example, when simulating by the digital computer the orbit trajectory of a spacecraft, say the space shuttle, improved estimates of the orbital parameters can be approximated numerically using the stratified Monte Carlo scheme given in Technical Report #4. Moreover the precision of the stratified scheme over an ordinary Monte Carlo simulation is significant. Thus this would allow the scientist or engineer to achieve a desired precision in the orbital parameter with many fewer samples from the computer. The procedure can be used either to reduce the variation in the orbital parameters calculated or to reduce the computer time necessary in order to achieve a desired precision.

Missing data problems occur in physical situations in many different ways, some of which are accidental and some of which are designed. The two technical reports on missing data were written under this grant covering those two major areas, i.e. accidental omissions and intentional ones. The monies saved by using partial data records (some records have elements missing) is considerable and is pointed out in some detail by Technical Report #6. Therein the economies are spelled out in terms of reduction in variation of the estimates, however these can immediately be translated into dollar figures given the various costs of observing each element of the observation vector. Similarly Technical Report #1 derives procedures for intentionally gathering incomplete observation vectors for reasons of economy. That is, one would optimally design an experiment to yield few observations on some variates and many observations on others, the choices of which depend not only on the cost per unit item but also the variation per unit item and the covariance between observations. Complete formulations of each of these procedures are available in the respective technical reports.

Finally several technical reports were written concerning the problems of estimating component of variance in an analysis of variance situation. Components of variance arise in linear models when there are several input variates (random variables) linearly combined to give an output variate. Various types of design are considered in these papers and a specific computer implementation of the technique so derived is also given.

IV. References

Harter, H. L., and A. H. Moore (1966). Local maximum likelihood estimation of the parameters of three-parameter lognormal populations from complete and censored samples. Journal of the American Statistical Association, 61, 842-851.

Hartley, H. O., and J. N. K. Rao (1967). Maximum likelihood estimation for the mixed analysis of variance model. Biometrika, 54, 93-108.

Hartley, H. O., M. George, and L. R. LaMotte (1969). Mixed convex and non-convex programming. Project Themis Technical Report # 21, Institute of Statistics, Texas A&M University.

Hartley, H. O., and R. R. Hocking (1971). Incomplete Data Analysis. Biometrics, 27, 783-825.

Hocking, R. R., and W. B. Smith (1968). Parameter estimation in the multivariate normal distribution with missing observations. Journal of the American Statistical Association, 63, 159-173.

Hocking, R. R., and W. B. Smith (1972). Optimum incomplete multinormal samples. Technometrics, 14, 299-309.

Ringer, L. J., and S. Suharto (1968). Stratified Monte Carlo sampling for symmetric statistics. Technical Report # 9, ARO Grant DA-ARO-31-124-G793, Institute of Statistics, Texas A&M University.

Syler, G. W. (1968). Estimation of parameters in the lognormal distribution with censored sampled. Unpublished Thesis, Texas A&M University.

Tiku, M. L. (1968). Estimating the parameters of lognormal distribution from censored samples, Journal of the American Statistical Association, 63, 134-140.