TREES - A COMPUTER PROGRAM FOR COMPLEX SURVEYS

D.T. Rylett, Health & Welfare Canada, D.R. Bellhouse, U. of Western Ontario
D.T. Rylett, Division of Biometrics, LCDC, Ottawa, Ontario, Canada K1A 0L2

KEY WORDS: multistage sampling design, endorder tree traversal, unbiased estimates, variance-covariance matrix

ABSTRACT

A computer package for the IBM PC has been developed to compute estimates of population means, proportions and totals and the associated variance-covariance matrix for a variety of estimators and designs for both single and multistage surveys. The underlying computer algorithm for the package involves an endorder traversal of a tree structure in which each stage of the tree corresponds to a stage in the sampling design and the nodes correspond to the sampling units from the previous stage. This type of algorithm is used to save disk space and to minimize computing time while obtaining exact estimates of the population variances and covariances.

INTRODUCTION

Bellhouse (1980, 1985) has provided computer algorithms for the estimation of the sampling variances of means, totals and proportions in complex surveys. These algorithms have been programmed and are now available in an interactive package program called TREES. The program runs on an IBM PC or compatible with or without a math co-processor.

Several computer programs are available for variance estimation in complex surveys. The most notable among them for the purposes of this paper is SUPERCARP, now available on a PC under the name of PC CARP. In some of the programs, for example CLUSTERS or SUPERCARP which are both described in Francis (1981), estimated standard errors or variances may be obtained for some specific sampling designs. In other programs, for example HES VAR X-TAB, described in Francis (1981), or subprograms in OSIRIS IV, described in Vinter (1980), the estimated variances for complex surveys are obtained by balanced repeated replication techniques. Thus, a survey researcher, when designing a survey in conjunction with these programs, is faced with one of two choices: choose a design which fits into one of the programs to obtain exact variance estimates, or choose a more general design and obtain approximate variance estimates. The computer program described in this paper is a generalization of the researcher's first choice. It provides a method to compute exact variance estimates for general complex sampling designs based on the associated finite population sampling theory. The program can also be easily used to calculate estimated variance components at each stage of a multistage survey so that the results can be used for planning purposes in subsequent surveys.

The program was originally developed by Bellhouse in 1979-80 on a PRIME mainframe. It contained variance estimation techniques for the estimation of totals in multistage designs using simple random sampling or pps sampling under Sampford's (1967) or randomized pps systematic sampling (see, for example, Hidiroglou and Gray, 1975). The code was poorly documented and the program was not user friendly. As an undergraduate computing project Briggs (1983), under the direction of Bellhouse, provided some documentation for the program. Rylett (1986), in a master's degree project done under the direction of Bellhouse, provided complete documentation to the program and a manual, expanded the program to include post-stratification through methods describe in Bellhouse (1985), and fixed the original code so that variance estimates for means and proportions would be available automatically. Currently, Rylett has transformed the program into a user-friendly package available interactively on an IBM PC or compatible personal computer.

SAMPLING THEORY

Consider a survey population consisting of N clusters from which a sample of n clusters is chosen. In the j-th cluster, totals on two covariables, x_j and y_j, may be obtained if the j-th cluster is chosen for the sample. A linear estimator of the population total Y, based on sample s, may be expressed as

$$\hat{Y} = \sum_{j \in s} w_j y_j$$

where w_j, for $j \in s$, are known weights fixed in advance or determined from population and sampled auxiliary variables. An expression for $\hat{X}$ is similarly obtained. The estimated covariance between $\hat{X}$ and $\hat{Y}$ may be described in general terms as cov ($\hat{X}$, $\hat{Y}$) = g($\mathbf{x_s}$, $\mathbf{y_s}$), a function of the sampled cluster totals, where $\mathbf{x_s}$ and $\mathbf{y_s}$ are the 1 X n vectors of sampled cluster totals, containing the elements x_j for $j \in s$ and y_j for $j \in s$ respectively. The estimated variance, var ($\hat{Y}$) = ($\mathbf{y_s}$, $\mathbf{y_s}$), is usually a quadratic form in $\mathbf{y_s}$. Rao and Vijayan (1977) have obtained the necessary form of the nonnegative quadratic unbiased estimate of the variance, var ($\hat{Y}$). The covariance can be obtained using the standard technique of finding the variance of $\hat{D} = \hat{X} - \hat{Y}$.

Two-stage sampling variances and covariances can be obtained using the unistage sampling formulae. In two-stage sampling the sampled cluster totals x_j and y_j are unknown but estimated at the second stage of sampling. On denoting the estimated cluster totals by $\hat{x}_j$ and $\hat{y}_j$, for $j \in s$, the 1 X n of estimated cluster totals may be denoted by $\hat{\mathbf{x}}_\mathbf{s}$ and $\hat{\mathbf{y}}_\mathbf{s}$ respectively. The estimated sampling covariance under two-stage sampling may be expressed as

$$\text{cov}\ (\hat{X}, \hat{Y}) = g\ (\hat{\mathbf{x}}_\mathbf{s}, \hat{\mathbf{y}}_\mathbf{s}) + \sum_{j\epsilon s} v_j\ c_j \quad (1)$$

based on estimates

$$\hat{X} = \sum_{j\epsilon s} w_j\ \hat{x}_j$$

and (2)

$$\hat{Y} = \sum_{j\epsilon s} w_j\ \hat{y}_j$$

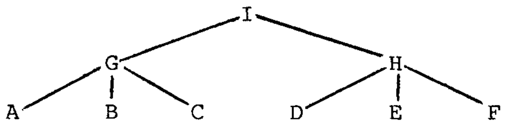
where w_j and v_j are known constants and c_j is the estimated covariance between x_j and y_j within the sampled primary $j \epsilon s$. The formula for stratified sampling is obtained upon setting $g(.,.) = 0$ in (1) and on taking the limit of summation in the remaining term in (1) as $j = 1,\ldots, n$.

Formulae (1) and (2) may be extended to three-stage and higher stages of sampling. However, for the purposes of numerical calculations, this extension is unnecessary. For any given multistage sampling design, the formulae may be used recursively to obtain numerical values of the estimates and estimated variances and covariances. Consider, for example, three-stage sampling; the extension to four or more stages is straightforward. In this situation, a sample of primary units is obtained, then samples of secondary units within each primary, and finally samples of tertiary units within each secondary. Begin at the final stage of sampling. Using the cluster sampling subroutines on the tertiary units, obtain estimates of the secondary totals or means and the associated variance-covariance estimates. Then go to the next stage up. Using formulae (1) and (2) with the estimates $\hat{\mathbf{x}}_\mathbf{s}$, $\hat{\mathbf{y}}_\mathbf{s}$ and $\hat{c}_i$ calculated from the previous stage, obtain estimates of the primary totals or means and the associated within primary variance-covariance estimates. Again, go to the next stage and repeat the same procedure. In this instance in formulae (1) and (2) $\hat{\mathbf{x}}_\mathbf{s}$ and $\hat{\mathbf{y}}_\mathbf{s}$ are the estimated primary totals and c_i, $i = 1,\ldots, n$, are the estimated covariances within primaries.

THREE STRUCTURES AND VARIANCE ESTIMATION

For any multistage survey carried out at a single point in time, a tree structure may be imposed on the sampled elements. On combining the sampling theory of Section 2 with tree traversal algorithms from computer science, variance estimates may be obtained which follow the structure of the sampling design.

The algorithm for the tree traversal follows the work of Bellhouse (1980). The method of picking the appropriate information to perform the necessary calculations is an endorder traversal. In this type of traversal, the subtrees are traversed from left to right with the root as the last node visited. Consider, for example, the tree presented in Figure 1.



Each letter correspond to a node on the tree. The traversal would start at A which is the leftmost node, and continue in the order ABCGDEFHI. A k-level tree can be imposed on a k-level sampling design. The k levels of sampling may include stratification and clustering. The sampling units at the (j - 1)th stage of the design correspond to the nodes of the j level of the tree, j = 2,..., k. The root of the tree (level 1) unifies the structure. The data level is the kth level of sampling while it is the (k+1)th level in the tree structure. Provided that the appropriate information is given at the nodes of the kth level of the tree, it is unnecessary to build the (k + 1)th level for the data. However, the data file must be appropriately ordered. For example, the tree in Figure 1 can be associated with a three-stage sampling design. The node labelled I unifies the tree; nodes G and H are associated with the primary sampling units; nodes A, B and C are associated with the secondary units within the primary "G", and nodes D, E and F are associated with the secondary units within the primary "H". To build the tertiary units, and hence the microdata or data file, into the tree structure, it is necessary only to store the number of sampled tertiary units at the nodes of the associated secondary units. In this case numbers n_A, n_B, ..., n_F would be stored in the nodes A, B, ..., F respectively. Since the traversal is endorder, the nodes will be visited in the given order. When node A is visited, the first n_A items or lines are read from the data file and the within secondary estimates and variance-covariance estimates are calculated from these data. When node B is visited, the n_B items at lines $n_A + 1$ to $n_A + n_B$ are read from the file and the appropriate estimates made. At node C the n_C items from the file are at lines $n_A + n_B + 1$ to $n_A + n_B + n_C$, and so on.

In general, when the tree is traversed, terminal and intermediate nodes can be reached. When a terminal node is reached, the data are picked out of the data file and are used to calculate the vector of estimates and the matrix of variance-covariance estimates. When an intermediate node is reached, the estimates from the subsample at that node are used to calculate the next level of estimates. When the root of the tree is attained, the overall survey estimates and variance-covariance estimates are calculated from the estimates of the previous stage. The endorder traversal works its way through each of the nodes of the stages to complete the calculations.

As an example of the calculations made at an intermediate node, consider again the tree in Figure 1, and the intermediate node G associated with a primary unit. Because of endorder traversal of the tree the nodes A, B, and C will have been visited prior to G. The calculations made at nodes A, B and C provide the quantities $\hat{x}_j$, $\hat{y}_j$ and $\hat{c}_j$ for $j \epsilon s$ used in formulae (1) and (2) to obtain the estimates and variance-covariance estimates within the primary associated with G. In general for an intermediate node will contain not only the values of the x's, y's and c's but also the additional information necessary to complete the calculations of formulae (1) and (2).

This necessary information to obtain estimates and variance-covariance estimates is input to the program during the construction of the tree. A node at level j (j = 1,..., k) in the tree contains the value of the number of branches of the node's subtree. This is also the sample size at the jth level of sampling. In addition to the sample sizes the following information is stored in a node at the j^{th} level (j = 1,..., k): a keyword describing the sampling design at the j^{th} stage used to obtain the subsample given by the nodes in the subtree at level j + 1, a keyword to describe the estimator to be used, the size variables, if PPS sampling was used, and the population size of the subsample if finite population correction factors are to be employed.

The program has been expanded to handle post-stratification in a multi-stage design. The method of calculating post-stratified variance estimates is based on the theory of Williams (1962). Suppose L post-strata are constructed. Let y denote the measurement on a sampling unit in the data file. Construct L new variables by setting $y_h = y$ if the sampling unit is in the hth post-stratum, 0 otherwise, h = 1,..., L. Make one pass through the tree structure which defines the sampling design. During this pass, calculate an estimate of the population mean for each of the L data sets defined by the variables y_h, h = 1,..., L. The resulting estimate $\hat{\bar{Y}}_h$ is the estimate of the mean in the post-stratum h, h = 1,..., L. The post-stratified estimate is $\hat{\bar{Y}}_p = \sum_{k=1} W_h Y_h$, where W_h, h = 1,..., L are known stratum weights provided in advance. Now transform the original data points y by setting $x = y - \hat{\bar{Y}}_h$ if the sampling unit is the hth post-stratum, h = 1,..., L. Then make a second pass through the tree structure. On this pass, calculate the estimated variance of $\hat{X}$, the estimated total based on the data x. The resulting estimate, var$(\hat{X})$ will be var$(\hat{Y}_p)$, the post-stratified variance estimate of the estimated total $\hat{Y}_p = N\hat{\bar{Y}}_p$ for the data y, where N is the total population size. The estimated variance of $\hat{\bar{Y}}_p$, var$(\hat{\bar{Y}}_p)$ = var$(\hat{Y}_p)/N^2$. This method requires the two passes through the data and the tree structure. However, only one set of operations by the program user is necessary: provide the stratum weights and the key words and numbers which describe the sampling design, the sample sizes, and other relevant information to perform the calculation.

PC VERSION OF TREES

The PC Version of TREES runs on an IBM or IBM compatible personal computer running on DOS 3.0 or later. It requires either a colour or monochrome monitor with CGA/EGA/VGA capabilites. TREES has been compiled so that it will utilize a math coprocessor if it is present but it is not necessary. The program can be run on a hard disk, a 360 Kb 5 1/4 inch floppy diskette or a 720 Kb 3 1/2 inch diskette. The executable code requires less than 200 K of disk space. The source code of the program consists of approximately 170 subroutines that were compiled using the Microsoft FORTRAN Optimizing Compiler Verison 4.01. When converting from the PRIME mainframe version to the PC, several subroutines were rewritten to produce smaller more efficient code. An emphasis was placed on writing modular code which would accomodate modification of the existing source code in later revisions. For example, if another probability proportional to size (PPS) sampling design were to be added to TREES the user would have to first write a FORTRAN subroutine which calculates the proper joint inclusion probabilites. Then, the subroutine which determines the PPS design to be used in the existing TREES source code must be changed to reflect the new option. Finally, the new and old source code must be recompiled and linked together to produce the required program. The main modification for the PC version of TREES is the improved input capability of the program. TREES now has interactive screens which prompt the user for the necessary information to built the sampling design (ie. number of stages, estimators, population and sample sizes, etc.), get the data and determine the appropriate calculations for the population estimates. TREES creates a command file from these responses which can be directly read into the program on subsequent runs. This command file can also be created by using a text editor and following a prearranged syntax. When post-stratifying a survey, the post-stratification variable now can be either continuous or discrete. For the continuous variable, the user must provide the cutpoints which define the post-stratum bounds. If required, the covariances between the population estimates can be calculated. The output of the program has been improved to include more information about the calculations that were performed. Also, the population estimates at all the stages in the survey can be produced if desired. Finally, an option also has been included to send the output from the program either to the screen or to a file on disk. This file then can be printed out and examined at at later time.

REFERENCES

Bellhouse, D.R. (1980): Computation of Variance-Covariance Estimates for General Multistage Sample Designs. COMPSTAT 1980: Proceedings in Computational Statistics, 57-63, Physica-Verlag, Vienna.

Bellhouse, D.R. (1985): Computing Methods for Variance Estimation in Complex Surveys. Journal of Official Statistics, 1 : 323-329.

Briggs, N.C. (1983): Computing Variance Estimates in Complex Surveys. Computer Science Project Report, McMaster University.

Francis, I. (1981): Statistical Software: A Comparative Review. North Holland, Amsterdam.

Hidiroglou, M.A. and Gray, G.B. (1975): A Computer Algorithm for Joint Probabilities of Selection. Survey Methodology, 1 : 99-108.

Rao, J.N.K. and Vijayan, K. (1977): On Estimating the Variance in Sampling with Probability Proportional to Aggregate Size. Journal of the American Statistical Association, 72 : 579-584.

Rylett, D.T. (1986): Variance-Covariance Estimation in Complex Surveys. Master's Project Report, University of Western Ontario.

Sampford, M.R. (1967): On Sampling without Replacement with Unequal Probabilities of Selection. Biometrika, 54 : 499-513.

Vinter, S. (1980): Survey Sampling Errors with OSIRIS IV. COMPSTAT 1980: Proceedings in Computational Statistics : 72-80, Physica-Verlag, Vienna.

Williams, W.H. (1962): The Variance of an Estimator with Post-Stratified Weighting. Journal of the American Statistical Association, 57 : 522-627.