Journal of Official Statistics
Vol. 1, No. 3, 1985, pp. 323–329

Computing Methods for Variance Estimation in Complex Surveys

D. R. Bellhouse[1]

Abstract: A description is given of a computer program which calculates estimates of the variance – covariance matrix for estimates of means, totals and proportions at any stage of a multistage sampling design. The computer program uses tree traversal algorithms in which the sampling design structure is made equivalent to an unbalanced tree. Extensions to post-stratification and variance estimation for complex statistics are also discussed.

Key words: Multistage sampling; variance estimation; tree structures.

1. Introduction

Several computer programs have been developed to estimate standard errors of population estimates in sample surveys. Francis (1981) has given a summary of eleven of these programs. 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 computational technique 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 computer program which implements these computational techniques is currently under development. To use this program, it is necessary only to provide the following information: the name of the sampling design and estimation procedure to be used at each stage, the size variable if a probability proportional to size sampling design has been used, the sample and population sizes, and the sample data in the appropriate order. The original method was described by Bellhouse (1980). A summary of this method is provided as well as extensions to post-stratification, to estimation of regression and other complex statistics, and to collapsed strata.

[1] Department of Statistical & Actuarial Sciences, University of Western Ontario, London, Ontario, Canada.

2. Variance-Covariance Estimation in General Multistage Designs

2.1. Sampling Theory Background

Consider a single-stage cluster sample of size n with sampled cluster totals $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ for two variables x and y. A linear estimator of the population total Y, of the variable y say, is $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ where w_i, $i = 1, \ldots, n$, are weights either fixed in advance or determined from population and sampled auxiliary variables. The estimated covariance between $\hat{X}$ and $\hat{Y}$ may be described in general terms as $\text{cov}(\hat{X}, \hat{Y}) = g(\underset{\sim}{x}_s, \underset{\sim}{y}_s)$, a function of the sampled cluster totals, where $\underset{\sim}{x}_s = (x_1, \ldots, x_n)$, $\underset{\sim}{y}_s = (y_1, \ldots, y_n)$ and s denotes the sample. The estimated variance, $\text{var}(\hat{Y}) = g(\underset{\sim}{y}_s, \underset{\sim}{y}_s)$, is usually a quadratic form in $\underset{\sim}{y}_s$. Rao and Vijayan (1977) have obtained the necessary form of the nonnegative quadratic unbiased estimate of the variance, $\text{var}(\hat{Y})$. The covariance can be obtained by the standard technique of finding the variance of $\hat{D} = \hat{X} - \hat{Y}$.

Most of the standard unistage sampling estimators are linear in y. Their variances and covariances can be obtained from sampling texts such as Cochran (1977) or derived from the result of Rao and Vijayan (1977). The first step in the development of the computer program was to write FORTRAN subroutines which obtained estimates of means or totals and variance-covariance estimates for various unistage sampling designs and estimators. These subroutines include: simple random sampling using the sample mean or ratio estimator; sampling with probability proportional to size (pps) using the Horvitz-Thompson estimator with Sampford's (1967) design or the randomized pps systematic sampling design with joint inclusion probabilities given by Hidiroglou and Gray (1975); and cluster sampling using simple random sampling of clusters with either the unbiased estimator or the ratio estimator, or using probability proportional to the size of the cluster. The subroutine for unistage ratio estimation may be used for calculating the separate ratio estimator and its variance estimate. Theoretically, any pps sampling design could be used in this program in conjunction with the Horvitz-Thompson estimator. It is necessary only for the program user to write a subroutine which calculates the joint inclusion probabilities for the given sampling design.

Two-stage sampling variances and covariances may be obtained using the unistage subroutines. Raj (1966) and Rao (1976) have obtained general formulations of the variance of $\hat{Y}$ where the estimate $\hat{Y}$ is based on a two-stage sample. Bellhouse (1980) has given the associated covariances for each method. Both methods are of the form

$$\text{cov}(\hat{X}, \hat{Y}) = g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s) + \sum_{i=1}^{n} v_i \hat{c}_i , \qquad (1)$$

based on estimates

$$\hat{X} = \sum_{i=1}^{n} w_i \hat{x}_i , \qquad (2)$$

and

$$\hat{Y} = \sum_{i=1}^{n} w_i \hat{y}_i ,$$

where $g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s)$ is a copy of $g(\underset{\sim}{x}_s, \underset{\sim}{y}_s)$ with $\underset{\sim}{x}_s$ replaced by $\hat{\underset{\sim}{x}}_s$ and $\underset{\sim}{y}_s$ replaced by $\hat{\underset{\sim}{y}}_s$. The coefficients w_i and v_i, $i = 1, \ldots, n$, are known constants, and $\hat{c}_i$, is the estimated covariance between $\hat{x}_i$ and $\hat{y}_i$ within the sampled primary i, $i = 1, \ldots, n$. Stratified sampling is obtained upon setting $g(\hat{\underset{\sim}{x}}_s, \hat{\underset{\sim}{y}}_s) = 0$ in (1) and $n = N$ in the remaining term of (1) where N is the population size of primaries or the total number of strata.

This general formulation can be used recursively to obtain estimates and variance-covariance estimates for any multistage design. Consider 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 $\hat{x}_s$, $\hat{y}_s$ 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) $\hat{x}_s$ and $\hat{y}_s$ are the estimated primary totals and $\hat{c}_i$, $i = 1, \ldots, n$, are the estimated covariances within primaries.

One way to computerize this general estimation procedure is to impose a tree structure on the sampling design. In the traversal of the tree, all the appropriate calculations are made.

2.2. *Tree Structures and Multistage Designs*

The terminology used here for tree structures is that of Knuth (1968). For a k-stage sampling design a k-level tree is constructed; and for a k-stage sampling design with stratification a$(k+1)$-level tree is constructed. The tree will be unbalanced if there are unequal sample sizes at any stage of sampling other than the final. The nodes at the ith level in the tree contain the relevant information about the ith stage of sampling. The number of nodes at the ith level of the tree corresponds to the number of sampling units at the $(i-1)$th stage of sampling. Measurements on the sampled units at the final stage of sampling are stored in a separate data file appropriately ordered.

The method is illustrated by a simple example. Consider Data Set 2 given by Kaplan et al. (1979) to test the accuracy of the calculations performed by a number of sample survey package programs. The design used was stratified two-stage sampling with three strata. Within each stratum, a two-stage sample of three primaries was chosen with five units within each primary. The tree structure which corresponds to this sampling design is given in Fig. 1. Below the tree in Fig. 1 is the data file appropriately ordered. The vertical lines below the data values indicate the boundaries of the subsamples at the final stage of sampling. The tree in Fig. 1 is a balanced tree of three levels containing thirteen nodes labelled A, B_i $(i = 1, 2, 3)$ and C_{ij} $(i = 1, 2, 3; j = 1, 2, 3)$. Node A is the root of the tree and nodes C_{ij} are terminal nodes.

The general tree construction algorithm used here has the following pattern. At any level in the tree, work from left to right. For each node in the current level, specify the number of nodes emanating from it to the next level. New storage locations for these lower level nodes and pointers to them are constructed. Then move to the next lower level. To construct the tree in Fig. 1, the number 3 is given to node A resulting in the creation of three storage locations for B_1, B_2, and B_3. Pointers to these storage locations are stored in A. The three branches from A to B_1, B_2, and B_3 correspond to the three strata in the design. The next step in the tree construction is to assign the number 3 to node B_1. Three new storage locations C_{11}, C_{12} and C_{13} are created and pointers to these locations are stored in B_1. The three branches in this subtree correspond to the three primary units chosen in stratum 1. Next, the number 3 is given to B_2 creating C_{21}, C_{22}, and C_{23} with the appropriate pointers in B_2. Finally, the number 3 is given to B_3 creating C_{31}, C_{32}, and C_{33}. At each step of the tree construction, additional information concerning the sampling design and desired estimator are given. At the root, node A, it is necessary to specify that the branches are strata. In each node $B_i(i = 1, 2, 3)$ the information given is that the design is simple random sampling of three primary units from a total of fifteen with the sample mean as the estimator. Finally, at

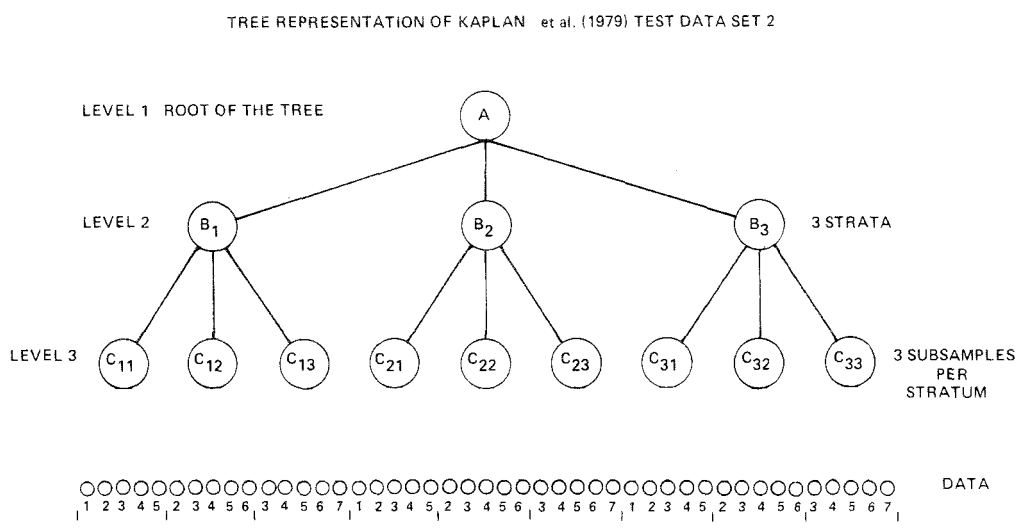

Fig. 1

the third level of the tree, the terminal nodes, C_{ij}, $i = 1, 2, 3$; $j = 1, 2, 3$ contain information about the "last stage of sampling," the subsampling of the primaries within the strata. In particular, the information given in each node is that the design is simple random sampling of five out of ten secondaries with the sample mean as the estimator. Finally, the data file that appears in Fig. 1 is in a specific order. The first five items belong to node C_{11}, i.e. they are the measurements on the five secondaries within the first primary in the first stratum. The next five items belong to node C_{12}, the following five to C_{13} and so on. Note that the data file is not part of the tree structure.

The type of traversal used in the program is endorder: the left-most subtree is traversed, then the second to left-most, and so on until the right-most subtree is traversed, and then the root of the subtree is visited. The algorithm used is a variation of one in Knuth (1968, pp. 317–319 and 560). At the time of construction of the tree, only the forward links exist. A pass could be made down the tree but not up. The backward links are created at the traversal stage. When a new node is reached in the traversal of the tree, a back pointer is given to its root. With endorder traversal, the nodes in Fig. 1 would be visited in the order $C_{11}C_{12}C_{13}B_1C_{21}C_{22}C_{23}B_2C_{31}C_{32}C_{33}B_3A$.

Assume that the population total is the item of interest. The calculations are performed as follows. Since the tree traversal is endorder, then the first node visited is C_{11}. The sample size of five indicates that five data points must be read from the data file, in this case the measurements 1, 2, 3, 4, 5. Using the information about the design and the estimate, a subroutine is called to calculate the estimate $\hat{Y} = 30$ and $v(\hat{Y}) = 25$. These values are stored in node B_1. The next node visited is C_{12} and the next five data items 2, 3, 4, 5, 6, are obtained from the file. The estimate $\hat{Y} = 40$ and $v(\hat{Y}) = 25$ are calculated from these data and stored in B_1. Then C_{13} is visited, data points 3, 4, 5, 6, 7 are read from the file and $\hat{Y} = 50$ and $v(\hat{Y}) = 25$ are stored in node B_1. Then node B_1 is visited. From the lower level of the tree or the lower stage in the design all the necessary data has been obtained. The estimate of the total for this primary is $\hat{Y} = 15(30 + 40 + 50)/3 = 600$ and the variance estimate

$v(\hat{Y}) = 15(15-3)[(30-40)^2 + (40-40)^2 + (50-40)^2]/[(3)\ (2)] + 15(25 + 25 + 25)/3 = 6375$. The values $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are stored in node A. The next nodes visited are C_{21}, C_{22} and C_{23}, in that order, from which values $\hat{Y} = 30, 40$ and 50 respectively and $v(\hat{Y}) = 25, 25$ and 25 respectively are calculated and stored in B_2. Then B_2 is visited and the calculations $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are made and stored in A. Then nodes C_{31}, C_{32} and C_{33}, in that order, are visited and $\hat{Y} = 30$, 40 and 50 respectively and $v(\hat{Y}) = 25$, 25 and 25 respectively are stored in node B_3. Then B_3 is visited and the calculation $\hat{Y} = 600$ and $v(\hat{Y}) = 6375$ are stored in A. Finally, node A is visited and $\hat{Y} = 600 + 600 + 600 = 1800$ and $v(\hat{Y}) = 6375 + 6375 + 6375 = 19125$ are calculated. The standard error of the estimate of the mean $(19125)^{1/2}/450 = .31$ agrees with the Kaplan et al. (1979) value.

The previous example is very simple and not indicative of typical survey data. Although the calculations for the preceeding example took only 3 seconds of CPU time on a PRIME 400 minicomputer, it remains to be seen whether the computing time would be excessive for larger and more complex surveys. Therefore the author obtained a larger data set which used a complex design. The data are from a survey of North American Indian children carried out in Canada during 1981–82. Six hundred responses with five variables each were analyzed. The Indians were divided into six strata by region of dwelling within Canada. Within a stratum, two, three or four enumeration areas (a Statistics Canada Census geographical area) were chosen by probability proportional to the size of the enumeration area. Sampford's (1967) design was assumed in the calculation of the joint inclusion probabilities. Within a chosen enumeration area, a number of families were chosen by simple random sampling and each child in a family was interviewed. The program produced both the estimates and estimated variance-covariance matrix. The calculations took 17 seconds of CPU time. The program also calculated and printed the estimates and variance-covariance estimates within each stratum so that interstratum comparisons could be made.

3. Post-Stratification

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, $h = 1, \ldots, L$. 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, $h = 1, \ldots, L$. The resulting estimate $\hat{\bar{Y}}_h$ is the estimate of the mean in the post-stratum h, $h = 1, \ldots, L$. The post-stratified estimate is $\hat{\bar{Y}}_p = \sum_{k=1}^{L} W_h \hat{\bar{Y}}_h$, where W_h, $h = 1, \ldots, L$ 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, $h = 1, \ldots, L$. 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, $\text{var}(\hat{X})$ will be $\text{var}(\hat{\bar{Y}}_p)$, 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 $\hat{\bar{Y}}_p$, $\text{var}(\hat{\bar{Y}}_p) = \text{var}(\hat{Y}_p)/N^2$.

This method requires 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.

4. Variance Estimates for the Simple Linear Regression Coefficient

The population regression coefficient may be expressed as

$$\hat{B} = \frac{\sum_{j=1}^{N} y_j(x_j - \bar{X})}{\sum_{j=1}^{N} (x_j - \bar{X})^2} ,$$

where N is the population size and the subscript j refers to an individual observation. This may be estimated by

$$\hat{B} = \frac{\sum_{j \in s} w_j z_{1j}}{\sum_{j \in s} w_j z_{2j}} , \qquad (3)$$

where s denotes the sample, unistage or multistage, w_j are the weights fixed in advance depending on the design, and where $z_{1j} = y_j(x_j - \hat{\bar{X}})$, $z_{2j} = (x_j - \hat{\bar{X}})^2$ and $\hat{\bar{X}} = \sum_{j \in s} w_j x_j / \sum_{j \in s} w_j$. Let the transformed variable

$$t_j = z_{1j} - \hat{B} z_{2j} . \qquad (4)$$

Three passes through the data and tree structure are necessary to calculate the estimated variance of $\hat{B}$. On the first pass, $\sum_{j \in s} w_j x_j$ and $\sum_{j \in s} w_j$, respectively, the estimated total for the x's and the estimated total for data which all have value 1, are calculated. After the second pass, $\hat{B}$ is calculated from (3). On this pass, both x and y are read from the data file and new variables z_1 and z_2 are derived. A one-pass algorithm could be derived to replace the first two passes. This would be analogous to the calculator and original formulae for sums of squares of deviations from a mean. As in this latter case, some numerical accuracy could be lost in the one-pass formulae. On the third pass through the tree structure and data file, calculate the estimated variance of $\hat{T}$, the estimated total based on the variable t from (4). Then $\text{var}(\hat{T})/(\sum_{j \in s} w_j z_{2j})^2$ is the required variance estimate, where $\text{var}(\hat{T})$ is the variance estimator based on the derived variable t.

The program could also be adapted to compute variance estimates for other nonlinear statistics provided that the estimates can be expressed as functions of estimated totals. For example, both the separate and combined ratio estimators and their variance estimates can be obtained in one pass through the data and tree structure. For the combined ratio estimator, say $\hat{R}_c$, estimates $\hat{\bar{X}}$, $\hat{\bar{Y}}$, $\text{var}(\hat{\bar{X}})$, $\text{var}(\hat{\bar{Y}})$ and $\text{cov}(\hat{\bar{X}}, \hat{\bar{Y}})$ can be obtained in one pass. Then $\hat{R}_c = \hat{\bar{Y}}/\hat{\bar{X}}$ and $\text{var}(\hat{R}_c) = \{\text{var}(\hat{\bar{Y}}) - 2\hat{R}_c \text{cov}(\hat{\bar{X}}, \hat{\bar{Y}}) + \hat{R}_c^2 \text{var}(\hat{\bar{X}})\}/\bar{X}^2$. For the separate ratio estimator of the total, estimates of the subtotals and their estimated variances at the last stage of sampling are obtained by ratio estimation using the appropriate unistage subroutine. The tree traversal proceeds in the usual manner using these estimates of totals and variance estimates. Rao (1982) has given a review of variance estimation techniques for ratios, multiple regression and correlation coefficients based on the Taylor linearization method.

5. Estimation with One Unit per Stratum

When stratification has been carried out to the extent that there is only one unit per stratum, the method given in Cochran (1977) or the method of Hansen, Hurwitz and Madow (1953) utilizing an auxiliary variable may be used in the program to obtain variance estimates. Only one pass through the tree structure and data file is necessary. During this pass, only estimates of means or totals are obtained at each stage. The final node visited in the tree by endorder traversal is the root of the tree. When this node is reached, estimates

of the stratum totals will have been calculated and stored in this node. With the size of the groups and the auxiliary variable, if present, specified beforehand, the variance estimates are obtained using formulae (5A.56) or (5A.57) in Cochran (1977).

6. References

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

Cochran, W.G. (1977): Sampling Techniques. 3rd Ed, Wiley, New York.

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

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953): Sample Survey Methods and Theory. Vol 1, Wiley, New York.

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

Kaplan, B., Francis, I. and Sedransk, J. (1979): Criteria for Comparing Programs for Computing Variances of Estimators from Complex Surveys: Proceedings of the 12th Interface Symposium, Waterloo, pp. 390–395.

Knuth, D.E. (1968): The Art of Computer Programming, Vol. 1. Addison-Wesley, Reading, Massachusetts.

Raj, D. (1966): Some Remarks on a Simple Procedure of Sampling without Replacement. Journal of the American Statistical Association, 61, pp. 391–396.

Rao, J.N.K. (1976): Unbiased Variance Estimation for Multistage Designs. Sankhya C, 37, pp. 133–139.

Rao, J.N.K. (1982): Some Aspects of Variance Estimation in Sample Surveys. Utilitas Mathematica, 21B, pp. 205–226.

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, pp. 579–584.

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

Vinter, S. (1980): Survey Sampling Errors with OSIRIS IV. COMPSTAT 1980: Proceedings in Computational Statistics, pp. 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, pp. 522–627.

Received September 1984
Revised May 1985