Journal of Official Statistics
Vol. 10. No. 4, 1994, pp. 395–405
© Statistics Sweden

A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys – Theory and Software Implementation

Claes Andersson[1] and Lennart Nordberg[1]

Abstract: This paper treats the estimation of standard errors in survey sampling. Let $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_J$ be linear (e.g., Horvitz-Thompson) estimators of the population totals $t_1, t_2, \ldots, t_J$. Simultaneous standard error estimation for a large number of functions $f_q(\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_J)$, $q = 1, 2, \ldots, Q$, is a common problem at statistical agencies. Such estimation can be very demanding even for a mainframe computer as the number of totals is large and the totals are allowed to represent completely arbitrary population domains.

We present a technique which reduces this problem to a manageable form and yields asymptotically unbiased variance estimates. This technique, which is based on Taylor approximation and an extension of the Woodruff transformation method, has recently been implemented in a computer program developed at Statistics Sweden. This program, called CLAN, can handle arbitrary rational functions of domain and population totals and auxiliary information can easily be included. CLAN was written in the SAS language and works on PCs as well as mainframes.

Key words: Survey sampling; variance estimation; statistical software.

1. Introduction

In sample surveys it is often desirable or necessary to employ estimates that are non-linear in the observations. Ratios, differences of ratios, regression coefficients and post-stratified means are common examples of such estimators. Another example is estimators that use auxiliary information in some way. Usually, exact expressions for the sampling variances of non-linear estimators are not available, neither are simple unbiased estimators of the variances.

Several different methods to compute estimates of these variances have been suggested (for an overview see Wolter 1985). The variances can be estimated by the use of a resampling plan, which leads to heavy computations, or by the use of some kind of random grouping. Another approach is to approximate the non-linear statistic with a linear function, i.e., by using the well known Taylor linearization method. The variance formula appropriate to the specific sampling design for the observed sample can then be applied to the approximation. This leads to a biased, but consistent estimator of the variance of the non-linear estimator.

[1] Statistics Sweden, S-70189 Örebro, Sweden.

The Taylor linearization technique has been used in survey sampling for a long time; examples of its applicability are given by Tepping (1968), Woodruff (1971) and Woodruff and Causey (1976). The technique has also been used in general computer programs for the estimation of ratios, regression coefficients, etc., and their variances. Examples of such computer programs are SUPER-CARP (Hidiroglou, Fuller, and Hickman 1976), SUDAAN (Research Triangle Institute 1989), PC-CARP (Schnell, Kennedy, Sullivan, Park, and Fuller 1988), and GES (Estevao, Hidiroglou, and Särndal 1995). Wolter (1985) provides an overview of available programs contemporary with the book.

2. A Computational Technique for Estimation of Standard Errors

2.1. *The problem*

Consider the problem of estimating a parameter θ of a fixed finite population U of size N. The parameter θ is a function of J totals $\mathbf{t} = (t_1, \ldots, t_j, \ldots, t_J)'$, that is

$$\theta = f(t_1, \ldots, t_j, \ldots, t_J) = f(\mathbf{t}) \qquad (2.1)$$

where $t_j = \Sigma_U y_{jk}$ is the total of the variable y_j in population U and y_{jk} is the value of y_j for unit k. A sample s is selected from U under the design $p(\cdot)$ and is used to calculate an estimate of t_j. A natural estimator of θ is to replace the different totals in $f(\mathbf{t})$ with their estimates $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J)'$, giving

$$\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J). \qquad (2.2)$$

We note that $\hat{\theta}$ is generally not an unbiased estimator of θ when $f(\mathbf{t})$ is non-linear, even when $\hat{\mathbf{t}}$ is an unbiased estimator of $\mathbf{t}$. It is, however, a consistent estimator of θ if $\hat{\mathbf{t}}$ is a consistent estimator of $\mathbf{t}$.

The approach discussed in the present paper cannot be used when the estimate $\hat{\theta}$ is defined implicitly via the solution of an estimating equation, for example, the median. In what follows we assume that the function f has an explicit form.

Another limitation is that our approach does not cover cases such as imputation for non-response where the estimation weights for non-responding units depend on the values for the responding units.

In most surveys, we are interested in estimates of different θs in specific subpopulations, called domains. Let the domain of interest be denoted U_j where $U_j \subset U$ and let c_j be an indicator variable such that

$$c_{jk} = \begin{cases} 1 \text{ if unit } k \in U_j \\ 0 \text{ otherwise.} \end{cases} \qquad (2.3)$$

Then we can do the transformation, $y_{jk} = y_k c_{jk}$. The total of a variable y in domain U_j is then $t_j = \Sigma_U y_{jk} = \Sigma_{U_j} y_k$.

In the following we will look upon $\mathbf{t}$ as a vector of totals generated by J arbitrary combinations of y-variables and domains. At one extreme $\mathbf{t}$ might be a vector of totals of *one variable* y in J different, possibly overlapping domains. At the other extreme $\mathbf{t}$ might be a vector of totals in *one domain* for J different y-variables.

2.2. *A solution*

The problem remains to find the design variance of $\hat{\theta}$ when $f(\mathbf{t})$ is a non-linear function. One approach, which will be adopted here, is to use the first term in the Taylor series approximation for the particular function f. The success of the method depends on the assumption that the sample size is so large that the higher order terms of the Taylor approximation can be neglected. In that case we may write

$$\hat{\theta} - \theta \approx \sum_{j=1}^{J} f_j'(\mathbf{t})(\hat{t}_j - t_j) \qquad (2.4)$$

where $f_j'(\mathbf{t}) = \partial f(\mathbf{t})/\partial t_j$ is the partial derivative of f evaluated at $\mathbf{t}$.

The mean squared error of $\hat{\theta}$, $MSE(\hat{\theta})$ is approximated by

$$MSE(\hat{\theta}) \approx V\left(\sum_{j=1}^{J} f_j'(\mathbf{t})\hat{t}_j\right) = \sum_{i=1}^{J}\sum_{j=1}^{J} f_i'(\mathbf{t})f_j'(\mathbf{t})\, C(\hat{t}_i, \hat{t}_j) \qquad (2.5)$$

where $C(\hat{t}_i, \hat{t}_j)$ is the covariance between $\hat{t}_i$ and $\hat{t}_j$. The sum contains $J(J+1)/2$ different variances and covariances that normally have to be estimated.

We note that the mean squared error of $\hat{\theta}$ is $MSE(\hat{\theta}) = V(\hat{\theta}) + Bias^2(\hat{\theta})$. Thus, if $Bias^2(\theta)$ is "small" compared to $V(\hat{\theta})$, as is usually the case when the sample size is "large", then $MSE(\hat{\theta})$ can be used as an approximation of $V(\hat{\theta})$. For a comprehensive discussion, see Wolter (1985).

The partial derivatives $f_1', \ldots, f_j', \ldots, f_J'$ usually depend on the unknown $\mathbf{t}$. For the purpose of variance estimation, we substitute sample based estimates of f_j' and $C(\cdot,\cdot)$. The f_j' is estimated by $f_j'(\hat{\mathbf{t}})$, i.e., the partial derivatives are evaluated at $\hat{\mathbf{t}}$ and $C(\cdot,\cdot)$ is estimated by $\hat{C}(\cdot,\cdot)$ computed from the sample. The result is the estimator

$$\hat{V}(\hat{\theta}) = \sum_{i=1}^{J}\sum_{j=1}^{J} f_i'(\hat{\mathbf{t}})f_j'(\hat{\mathbf{t}})\, \hat{C}(\hat{t}_i, \hat{t}_j). \qquad (2.6)$$

In general, $\hat{V}(\hat{\theta})$ will not be an unbiased estimator of either the true $MSE(\hat{\theta})$ or $V(\hat{\theta})$. It is however a consistent estimator of $MSE(\hat{\theta})$ provided that $\hat{t}_j$ and $\hat{C}(\hat{t}_i, \hat{t}_j)$ are consistent estimators of t_j and $C(\hat{t}_i, \hat{t}_j)$.

Suppose that the ordinary ratio estimator $\hat{t}_{\text{ratio}} = (t_x/\hat{t}_x)\hat{t}_y$ is used to estimate the total of y. Obviously $\hat{t}_{\text{ratio}}$ is an estimator for a total but in our approach it will be considered (and covered) as a *function* of the elementary estimators $\hat{t}_x$ and $\hat{t}_y$ of t_x and t_y and the known t_x. We will later state the exact form required of our elementary estimators, but for now we simply assume that they are linear, i.e., $\hat{t}_j = \Sigma_s w_k y_{jk}$, $k \in s$. It is then possible to reverse the order of summation and use the transformation

$$z_k = \sum_{j=1}^{J} f_j'(\mathbf{t}) y_{jk}. \qquad (2.7)$$

The total $t_z = \Sigma_U z_k$ is estimated by $\hat{t}_z = \Sigma_s w_k z_k$ and the variance $V(\hat{t}_z)$ is an approximation of $MSE(\hat{\theta})$. The estimation of $V(\hat{t}_z)$ has a well known solution, at least for sampling designs used in practice. Thus by the use of the single variable z which is a linear combination of the original variables $y_1, \ldots, y_J$ we have converted a J-variate estimation problem into a simple univariate estimation problem.

In order to estimate $MSE(\hat{\theta})$, we replace $\mathbf{t}$ with sample based estimates, resulting in the variable

$$\hat{z}_k = \sum_{j=1}^{J} f_j'(\hat{\mathbf{t}}) y_{jk}. \qquad (2.8)$$

This expression is due to Woodruff (1971). The variance of $\hat{t}_z$ is easily estimated for the appropriate design by using $\hat{z}_k$ and the ordinary formulas for estimating the variance of a total.

In most surveys, we are not interested in just one function $f(\mathbf{t})$, but in several functions, maybe hundreds, each one with its own $\mathbf{t}$, $\theta_q = f_q(\mathbf{t}_q)$, $q = 1, \ldots, Q$, where $\mathbf{t}_q$ contains J_q totals. The obvious solution is to substitute sample based estimates of $\mathbf{t}_q$ in f_q.

Organising the calculation of the variances of $\hat{\theta}_q$, $q = 1, \ldots, Q$, based on the estimator (2.6) can be a very demanding task even with a mainframe computer. Two major problems arise:

i. We must keep track of all pairs of totals and compute their covariances. The number of pairs is $\sum_{q=1}^{Q} J_q(J_q - 1)/2$ which can be very large in many common applications. The problem is magnified if we allow arbitrary combinations of domains and y-variables. This means that a sample unit may contribute to many domains, totals and functions.
ii. We must provide partial derivatives of all functions allowed.

The first problem is reduced by using Woodruff's transformation (2.8). The use of Woodruff's transformation for an arbitrary function has been limited by the computational difficulty of evaluating the required partial derivatives, problem (ii) above. In order to solve this problem, Woodruff and Causey (1976) suggested a method for the numerical estimation of first derivatives. However, instead of using an approximate and cumbersome numerical method, we are able to evaluate partial derivatives from the expression of $f(\mathbf{t})$ alone and by applying (2.8) in a stepwise fashion.

Using the transformation (2.8) has one slight drawback. In order to compute the z-values we need the estimates $\hat{t}_q$ for each q. That means we have to make two passes through the data. In the first pass we compute the estimates $\hat{t}_q$, in the second pass we compute $\hat{\theta}_q$, the z-values and $\hat{V}(\hat{t}_{zq}) = \hat{V}(\hat{\theta}_q)$ for each q. However, compared to the problem (i) above we consider this a minor problem.

Problem (ii) can be solved in the following manner. Let $f(\mathbf{t})$ be a compound function which can be written in the following form

$$f(\mathbf{t}) = G(g_1(\mathbf{t}), g_2(\mathbf{t})) \tag{2.9}$$

where g_1, g_2 and G are sufficiently differentiable functions.

Now we can apply Woodruff's transformation on $g_1(\mathbf{t})$ and $g_2(\mathbf{t})$, giving

$$z_{1k} = \sum_{j=1}^{J} \frac{\partial g_1(\mathbf{t})}{\partial t_j} y_{jk} \quad \text{and}$$

$$z_{2k} = \sum_{j=1}^{J} \frac{\partial g_2(\mathbf{t})}{\partial t_j} y_{jk}. \tag{2.10}$$

The transformation of f is given by (2.7). Using the chain-rule we get

$$f_j'(\mathbf{t}) = \frac{\partial f(\mathbf{t})}{\partial t_j} = \frac{\partial G}{\partial g_1}\frac{\partial g_1}{\partial t_j} + \frac{\partial G}{\partial g_2}\frac{\partial g_2}{\partial t_j}. \tag{2.11}$$

Inserting (2.10) and (2.11) into (2.7) we get

$$z_k = \frac{\partial G}{\partial g_1} z_{1k} + \frac{\partial G}{\partial g_2} z_{2k}. \tag{2.12}$$

Thus it is possible to obtain the transformation (2.7) in a stepwise fashion.

2.3. *An example*

Let us illustrate the technique with an example. Suppose we want to estimate the variance of the function

$$\hat{\theta} = \frac{\hat{t}_1 \cdot \hat{t}_2}{\hat{t}_3 \cdot \hat{t}_4}. \tag{2.13}$$

Under (2.7), taking the partial derivatives of $\hat{\theta}$, and some algebraic manipulations yield

$$\hat{z}_k = \frac{\hat{t}_1 \cdot \hat{t}_2}{\hat{t}_3 \cdot \hat{t}_4}\left(\frac{y_{1k}}{\hat{t}_1} + \frac{y_{2k}}{\hat{t}_2} - \frac{y_{3k}}{\hat{t}_3} - \frac{y_{4k}}{\hat{t}_4}\right). \tag{2.14}$$

However, we can get exactly the same expression by applying the transformation in two steps. Let

$$\hat{\theta}_1 = g_1(\hat{\mathbf{t}}) = \frac{\hat{t}_1}{\hat{t}_3} \tag{2.15}$$

and

$$\hat{\theta}_2 = g_2(\hat{\mathbf{t}}) = \frac{\hat{t}_2}{\hat{t}_4} \tag{2.16}$$

then

$$\hat{\theta} = \hat{\theta}_1 \cdot \hat{\theta}_2. \tag{2.17}$$

By (2.10) we get

$$\hat{z}_{1k} = \frac{\hat{t}_1}{\hat{t}_3}\left(\frac{y_{1k}}{\hat{t}_1} - \frac{y_{3k}}{\hat{t}_3}\right) \quad \text{and}$$

$$\hat{z}_{2k} = \frac{\hat{t}_2}{\hat{t}_4}\left(\frac{y_{2k}}{\hat{t}_2} - \frac{y_{4k}}{\hat{t}_4}\right). \tag{2.18}$$

Inserting (2.18) into (2.12) yields

$$\hat{z}_k = \hat{\theta}_1 \cdot \hat{\theta}_2\left(\frac{\hat{z}_{1k}}{\hat{\theta}_1} + \frac{\hat{z}_{2k}}{\hat{\theta}_2}\right). \tag{2.19}$$

It is easy to see that (2.19) is the same as (2.14). Note that (2.19) is the transformation we would get from the product $\hat{\theta}_1 \cdot \hat{\theta}_2$ using $\hat{z}_{1k}$ and $\hat{z}_{2k}$ as input data.

It would also have been possible to define $\hat{\theta}_1$ and $\hat{\theta}_2$ in a different way and still get exactly the same result. For example, define $\hat{\theta}_1 = \hat{t}_1 \cdot \hat{t}_2$ and $\hat{\theta}_2 = \hat{t}_3 \cdot \hat{t}_4$ then $\hat{\theta} = \hat{\theta}_1/\hat{\theta}_2$ and the appropriate expressions for the z-variables will give the same result as (2.14).

3. Software Implementation

3.1. General remarks

Since it is possible to compute the derivatives of $\hat{\theta}$ in a stepwise manner, it is relatively easy to construct an algorithm that is suitable for a computer program where the user does not have to bother with these derivatives. If we allow only rational functions of totals it means that we only have to worry about the derivatives of functions like $t_1 \, op \, t_2$, where *op* is one of the operators $+$, $-$, $\times$ and $/$.

Note: Although it is rather simple to include other functions, i.e., $\theta = func(t_0)$, where *func* is, for example, $log(t_0)$, $exp(t_0)$, $sqrt(t_0)$, etc., we do not treat these functions here.

The following table shows the well known Woodruff or z-transformations needed to estimate the variance of $\theta = t_1 \, op \, t_2$

Table 1. z-transformations for different operators

op	*z-transformation*
$+$	$z_k = y_{1k} + y_{2k}$
$-$	$z_k = y_{1k} - y_{2k}$
$\times$	$z_k = \theta \times (y_{1k}/t_1 + y_{2k}/t_2)$
$/$	$z_k = \theta \times (y_{1k}/t_1 - y_{2k}/t_2)$

In order to compute the variance of an arbitrary rational function of totals we propose a two-pass algorithm, which means that one has to make two passes through the data set. In the first pass all the totals needed are estimated, for example, by the Horvitz-Thompson estimator $\hat{t}_j = \Sigma_s y_{jk}/\pi_k$ where y_{jk} is defined as before and π_k is the inclusion probability of unit k.

When all totals are estimated, the second pass begins. The parameters θ_q of interest may be specified in at least two ways, A and B.

In approach A the user specifies the parameter(s) successively by using totals and intermediate transformations in a pairwise manner. In B the user specifies the parameter(s) of interest in terms of all totals involved and lets the computer do the pairwise decomposition.

We illustrate the two approaches A and B by a simple example. Let the parameter of interest be the same as in Section 2, i.e., we want to find the z-transformation needed to estimate the variance of $\hat{\theta} = \hat{t}_1 \cdot \hat{t}_2/\hat{t}_3 \cdot \hat{t}_4$.

The totals $t_1, \ldots, t_4$ are estimated according to the sample design used when the sample s was taken. In approach A the user is supposed to specify θ in three steps in this case. Note that $y_{jk} = 0$ if

Table 2. Intermediate and final z-transformations when a ratio between two products is of interest

step	*estimate*	*z-transformation*
1	$\hat{\theta}_1 = \hat{t}_1 \cdot \hat{t}_2$	$\hat{z}_{1k} = \hat{\theta}_1(y_{1k}/\hat{t}_1 + y_{2k}/\hat{t}_2)$
2	$\hat{\theta}_2 = \hat{\theta}_1/\hat{t}_3$	$\hat{z}_{2k} = \hat{\theta}_2(\hat{z}_{1k}/\hat{\theta}_1 - y_{3k}/\hat{t}_3)$
3	$\hat{\theta}_3 = \hat{\theta}_2/\hat{t}_4$	$\hat{z}_{3k} = \hat{\theta}_3/(\hat{z}_{2k}/\hat{\theta}_2 - y_{4k}/\hat{t}_4)$

unit k does not contribute to the estimate of t_j.

The variance of $\hat{\theta}_3$, i.e., of $\hat{\theta}$, is estimated by using $\hat{z}_{3k}$ and using the formula for the variance of $\hat{t}_z = \Sigma_s \hat{z}_{3k}/\pi_k$. The operations on $\hat{t}_1$, $\hat{t}_2$, $\hat{t}_3$ and $\hat{t}_4$ in steps 1–3 may be taken in different orders, all leading to the same result.

In approach B, the user specifies $\theta = (t_1 \cdot t_2)/(t_3 \cdot t_4)$, for example. The computer program analyses the expression, evaluates it and computes the appropriate z-values in a pairwise manner as above by using the usual precedence rules, first × and /, then + and −. Of course the precedence may be changed by parenthesis. The main difference between approaches A and B is that the computer does more of the work in B than in A. Next we describe a computer program that uses approach A.

3.2. CLAN

CLAN is a program designed to estimate standard errors in survey sampling. The program was written in the SAS language (SAS Institute, Inc. 1988). CLAN permits the user to choose between a large number of estimators, including estimators that use auxiliary information, and it can handle very complex combinations of domain specifications. The major strength of the program lies in the flexibility with which the user may combine estimators with the specification of complex sets of domains. This section contains an overview of CLAN. For more details about the user interface and various practical aspects, we refer to Andersson and Nordberg (1992).

Let $\mathbf{t} = (t_1, \ldots, t_j, \ldots, t_J)'$ be a vector of J population or subpopulation (domain) totals as defined in Section 1. Let the parameter of interest be

$$\theta = f(\mathbf{t}) \qquad (3.1)$$

where f is an arbitrary rational function. As an estimator of θ CLAN uses

$$\hat{\theta} = f(\hat{\mathbf{t}}) \qquad (3.2)$$

where $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J)'$ is a linear estimator of $\mathbf{t}$.

CLAN computes $\hat{\theta}$ and an estimate of the standard error $\sqrt{V(\hat{\theta})}$, using the technique described in Section 2. The output of CLAN contains one or more pair(s)

$$(\hat{\theta}, \sqrt{\hat{V}(\hat{\theta})}) \qquad (3.3)$$

which are given prefixes P for point estimates and S for standard error estimates, respectively.

In one step CLAN can do the calculations of (3.3) for many different sets of totals and domains and different functions. The computation is based on the technique referred to as approach A in Section 3.1.

The form of the function f and the definition of the totals in $\mathbf{t}$ must be supplied by the user. The estimator $\hat{\mathbf{t}}$ is specified by the user by choosing one of a number of available strategies, i.e., combinations of sampling design and estimator, including the choice of non-response model.

So far four strategies have been implemented in CLAN. The majority of surveys conducted at Statistics Sweden reduce to these four strategies, including a number of surveys that use pps-sampling, various types of network sampling and two-phase sampling schemes for stratification.

The point and standard error estimators used in the different strategies are found in the Appendix.

Strategy 1 assumes stratified sampling with simple random sampling without replacement within strata (SRS). The non-response model presumed is independent responses with equal response probabilities within strata.

Strategy 2 assumes the same design as in strategy 1. The difference is that strata can be divided into response groups with equal response probabilities within groups. Population group sizes are assumed unknown. This strategy also covers two-phase sampling.

Strategy 3 is used for one-stage cluster sampling where clusters can be divided into strata and selection is SRS without replacement within strata. All the elements of the selected clusters are included in the sample. The non-response model is the same as in strategy 1, i.e., the response probabilities of the clusters are assumed to be equal within strata.

Strategy 4 can be seen as a combination of strategies 2 and 3. The design is taken from strategy 3 and the non-response model from strategy 2 is assumed for the clusters.

It should perhaps be emphasised that CLAN allows different elements within the same cluster to belong to different domains.

The reader may have noticed that multistage cluster sampling is missing. The reason is that such designs are currently used very little at Statistics Sweden. It could be included as a separate strategy in the future if necessary. However, it is of course, possible to let CLAN compute *approximate* estimates of the variances in multistage designs by assuming strategy 3 and computing appropriately weighted estimates of the PSU totals.

Next we discuss the specification of the function f. CLAN is at present restricted to rational functions. Other types of functions could easily be added if necessary since the stepwise Woodruff technique applies to any sufficiently differentiable function.

If f is rational then the Woodruff transformation for f can be obtained by successive use of the Woodruff transformations corresponding to addition, subtraction, multiplication or division of *two* totals or functions of totals. These elementary transformations have been pre-programmed and are available to the user as the SAS macros: *%ADD*, *%SUB*, *%MULT* and *%DIV*.

Each SAS macro contains three parameters. For instance, *%SUB*(Z, $Z1$, $Z2$) obtains the Woodruff transformation for $Z = Z1 - Z2$.

We will illustrate how the macros work in an example. This example, among others, is worked out in more detail in Andersson and Nordberg (1992).

3.3. *Example*

Let T_{ab} and N_{ab} be "Total income" and "Number of individuals", respectively, in cell (a, b) in the following table, (index a represents social group and index b age group).

Remark: The social grouping is partly overlapping since workers are part of the group employees. One feature of CLAN is that domains can be arbitrarily overlapping.

Suppose that we want point and standard error estimates for "Mean income per individual," $R_{ab} = T_{ab}/N_{ab}$, and the relative mean income in cell (a, b) proportional to mean income in the whole age group b,

Table 3. A skeleton table for Age × Social group

	Age group			
Social group	–29	30–44	45–64	All
Farmers	·	·	·	·
Workers	·	·	·	·
Employees	·	·	·	·
All	·	·	·	·

$Q_{ab} = R_{ab} \cdot N_{4b}/T_{4b}$. Notice that row 4 contains "all social groups."

The function f is two-dimensional here, $f(a,b) = (R(a,b), Q(a,b))$. To specify f the user must write a *%macro FUNKTION*(a,b). This is not a difficult task for a user who is familiar with elementary SAS programming. CLAN provides a tool-kit which consists of six pre-programmed macros including the four already mentioned.

All the totals involved in the function, i.e., T_{ab}, N_{ab}, T_{4b} and N_{4b} must be defined for all (a,b). This is done by invoking a pre-programmed macro *%TOT*(*,*,*). For example, the total income in cell (a,b) is obtained by *%TOT*(TAB, INCOME, (SOCG = &*a*) AND (AGE = &*b*)) telling CLAN that we want the units that meet the condition (SOCG = &*a*) AND (AGE = &*b*) to contribute to the estimate of the total of variable INCOME and name the estimate "Tab." The variables INCOME, SOCG and AGE are assumed to exist in the input data set. The totals N_{ab}, T_{4b} and N_{4b} are defined analogously. (The SAS macro language uses "&*a*" to refer to the *value* of variable *a*.) The functions R and Q are then obtained as follows:

%DIV(*RAB*, *TAB*, *NAB*) $\quad \mathrm{R}_{ab} = T_{ab}/N_{ab}$

%DIV(*R*4*B*, *T*4*B*, *N*4*B*) $\quad \mathrm{R}_{4b} = T_{4b}/N_{4b}$

%DIV(*QAB*, *RAB*, *R*4*B*) $\quad \mathrm{Q}_{ab} = \mathrm{R}_{ab}/\mathrm{R}_{4b}$

This will provide the Woodruff transformation for every (a,b) for the functions R_{ab}, R_{4b} and Q_{ab}. (The function R_{4b} is not of primary interest here. It serves mainly as an intermediate result.)

An alternative way to obtain R_{ab} and Q_{ab} is, for example, as follows:

%DIV(*RAB*, *TAB*, *NAB*)

%MULT(*XAB*, *RAB*, *N*4*B*)

%DIV(*QAB*, *XAB*, *T*4*B*)

The user can construct any rational function or set of functions by using the elementary macros as building blocks in much the same way as one would use the operations addition, subtraction, multiplication and division in elementary algebra.

Finally we state the functions for which we want to compute point estimates and standard errors. (The operations above only provide the appropriately transformed data for each sample unit.)

By invoking a pre-programmed macro *%ESTIM* we get the requested point and standard error estimates. This is done here by the following commands.

%ESTIM(*RAB*)

%ESTIM(*QAB*)

It is sometimes useful to do a bit of elementary SAS programming when writing *%macro FUNKTION*. Example 2 on the post-stratified ratio estimator in Andersson and Nordberg (1992) illustrates this technique.

Appendix. Estimation Formulas Under Strategies 1 Through 4

The computation of variances for functions of totals is always converted into computation of variances of totals in CLAN through the use of the Woodruff transformation. When $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_j, \ldots, \hat{t}_J)'$ where $\hat{t}_j = \Sigma_s w_k y_{jk}$ has been calculated, then θ is estimated by $\hat{\theta} = f(\hat{\mathbf{t}})$ and the transformation $\hat{z}_k = \sum_{j=1}^{J} f_j'(\hat{\mathbf{t}}) y_{jk}$ is performed. The variance of $\hat{\theta}$ is then estimated by $\hat{V}(\hat{t}_z)$ where $\hat{t}_z = \Sigma_s w_k z_k$. Thus only formulas for point and variance estimation of totals are needed.

In this appendix we specify the formulas used under each of the four strategies allowed in CLAN.

A.1. Strategy 1

Let N_h be the number of units in stratum h in the sampling frame, $h = 1, 2, \ldots, H$. Let m_h be the number of responding sampling units i stratum h.

The total t of variable y is estimated by

$$\hat{t} = \sum_{h=1}^{H} \frac{N_h}{m_h} \sum_{k=1}^{m_h} y_{hk} \tag{A.1}$$

and the variance of $\hat{t}_z$ is estimated by

$$\hat{V}(\hat{t}_z) = \sum_{h=1}^{H} \frac{N_h^2}{m_h} \left(1 - \frac{m_h}{N_h}\right) \frac{1}{m_h - 1} \left[\sum_{k=1}^{m_h} z_{hk}^2 - \frac{\left(\sum_{k=1}^{m_h} z_{hk}\right)^2}{m_h} \right]. \tag{A.2}$$

A.2. Strategy 2

Stratum h is divided into L_h response homogeneity groups. The units in a given group are assumed to respond independently and with the same probability. The population size is assumed to be unknown in each group.

In stratum h we know

N_h the number of population units,

n_h the number of sampling units.

In response homogeneity group hg, $g = 1, 2, \ldots, L_h$, i.e., group g in stratum h we know

n_{hg} the number of sampling units in group hg,

m_{hg} the number of responding sampling units in group hg.

The total t of variable y is estimated by

$$\hat{t} = \sum_{h=1}^{H} \frac{N_h}{m_h} \sum_{g=1}^{L_h} \frac{n_{hg}}{m_{hg}} \sum_{k=1}^{m_{hg}} y_{hgk}. \tag{A.3}$$

We can also write (A.3) as follows

$$\hat{t} = \sum_{h=1}^{H} \sum_{g=1}^{L_h} \frac{\hat{N}_{hg}}{m_{hg}} \sum_{k=1}^{m_{hg}} y_{hgk} \tag{A.4}$$

where $\hat{N}_{hg}$ is an unbiased estimator of the unknown N_{hg}.

The variance of $\hat{t}_z$ is estimated by, (see Särndal, Swensson, and Wretman 1992, p. 582)

$$\hat{V}(\hat{t}_z) = \sum_{h=1}^{H} N_h(N_h - 1) \sum_{g=1}^{L_h} \frac{n_{hg}}{n_h} \left(\frac{n_{hg} - 1}{n_h - 1} - \frac{m_{hg} - 1}{N_h - 1}\right) \frac{1}{m_{hg}} s_{zhg}^2$$
$$+ \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h - 1} \sum_{g=1}^{L_h} \frac{n_{hg}}{n_h} (\bar{z}_{hg} - \bar{\bar{z}}_h)^2 \tag{A.5}$$

where

$$\bar{z}_{hg} = \frac{1}{m_{hg}} \sum_{k=1}^{m_{hg}} z_{hgk}, \quad \bar{\bar{z}}_h = \sum_{g=1}^{L_h} \frac{n_{hg}}{n_h} \bar{z}_{hg} \quad \text{and} \quad s_{zhg}^2 = \frac{1}{m_{hg} - 1} \left[\sum_{k=1}^{m_{hg}} z_{hgk}^2 - m_{hg} \bar{z}_{hg}^2 \right].$$

A.3. Strategies 3 and 4

These strategies are used for one-stage cluster sampling where clusters can be divided into strata and selected with SRS within strata. All the elements of the sampled clusters are included in the sample. Non-response of clusters under strategy 3 is treated in the same way as non-response of sampling units under strategy 1.

Strategy 4 corresponds to strategy 2 in the sense that clusters can be divided into response homogeneity groups with unknown population group sizes. The formulas (A.1)–(A.5) are still valid under strategies 3 and 4. Nevertheless, strategies 3 and 4 define the variables y and z differently, and are different in this respect.

In strategy 3, y_{hk} is defined by

$$y_{hk} = \sum_{\nu=1}^{M_{hk}} y_{hk\nu},$$

where M_{hk} is the number of elements in cluster hk and $y_{hk\nu}$ is the value of y for element ν in cluster hk.

In strategy 4, y_{hgk} is defined by

$$y_{hgk} = \sum_{\nu=1}^{M_{hgk}} y_{hgk\nu},$$

where M_{hgk} is the number of elements in cluster hgk and $y_{hgk\nu}$ is the value of y for element ν in cluster hgk.

In strategies 3 and 4 the z-values are calculated using y_{hk} and y_{hgk} as defined above. Notice that different elements in the same cluster may belong to different domains.

4. References

Andersson, C. and Nordberg, L. (1992). CLAN – A Program for the Computation of Standard Errors in Survey Sampling. Technical Report, Statistics Sweden.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, in press.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1976). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.

Research Triangle Institute (1989). SUDAAN: Professional Software for SUrvey DAta ANalysis, Version 5.3. Research Triangle Park, NC, U.S.A.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer Verlag.

SAS Institute, Inc. (1988). SAS Language Guide for Personal Computers; Release 6.03 Edition. Cary, NC: SAS Institute Inc.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, J.P., and Fuller, W.A. (1988). Personal Computer Variance Software for Complex Surveys. Survey Methodology, 14, 59–69.

Tepping, B. (1968). The Estimation of Variance in Complex Surveys. Proceedings of Social Statistics Section, American Statistical Association, 11–18.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer Verlag.

Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, 66, 411–414.

Woodruff, R.S. and Causey, B.D. (1976). Computerized Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, 71, 315–321.

Received November 1993
Revised July 1994