Ann. Inst. Statist. Math.
Vol. 53, No. 2, 404–417 (2001)

GENERALIZED CALIBRATION APPROACH FOR ESTIMATING VARIANCE IN SURVEY SAMPLING

SARJINDER SINGH*

Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada N9B 3P4

(Received January 27, 1999; revised July 21, 1999)

Abstract. In the present investigation, a general set-up for inference from survey data that covers the estimation of variance of estimators of totals and distribution functions has been considered, using known higher order moments of auxiliary information at the estimation stage. Several estimators of variance of estimators of totals and distribution functions are shown to be the special cases of the proposed strategy. An empirical study has also been given to show the performance of the proposed estimators over the existing estimators in the literature.

Key words and phrases: Auxiliary information, distribution functions, totals, estimation of variance.

1. Introduction

In survey sampling, known auxiliary information is often used at the estimation stage to increase the precision of the estimators of population variance. Das and Tripathi (1978), Srivastava and Jhajj (1980), Isaki (1983) and Garcia and Cebrian (1996) have considered the problem of estimation of finite population variance using known variance of the auxiliary information. Wu (1982), Wolter (1985), Deng and Wu (1987) and Särndal (1982) have considered the problems of estimation of variance of ratio and regression estimators of totals using known information of total of the auxiliary variable. Rao (1994) reported that several estimators of a population distribution function have also been proposed using auxiliary information at the estimation stage. Chaudhuri and Roy (1997) have suggested optimal variance estimation techniques for generalized regression predictor by assuming that population total of the auxiliary character is known. Singh *et al.* (1998) have proposed a higher order calibration approach to estimate the variance of the general linear regression estimator using known variance of the auxiliary information.

The main purpose of this paper is to provide a general set-up that can be used to estimate the variance of estimators of totals or distribution functions using known information about the second order moments of the auxiliary characteristic. The higher order calibration approach proposed by Singh *et al.* (1998) is also a special case of the proposed strategy.

*Now at Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK S7N 5E6 Canada.

2. General parameter: notations

Suppose a population Ω consists of N distinct units identified through the labels $j = 1, 2, \ldots, N$. A sample is a subset, s, of Ω and the associated y-values, i.e. $\{(i, y_i), i \in s\}$, selected according to a specified sampling design which assigns a known probability $p(s)$ to s such that $p(s) > 0$ for all $s \in S$, the set of possible samples s, and $\sum_{s \in S} p(s) = 1$. Following Rao (1994), we consider general parameters of interest:

$$H_y = \sum_{j \in \Omega} h(y_j) \quad \text{and} \quad \bar{H}_y = N^{-1} H_y \tag{2.1}$$

for a specified function h. The choice of $h(y) = y$ gives the population total $H_y = Y$ and the population mean $\bar{H} = \bar{Y}$, while the choice $h(y) = \Delta(t - y)$ with $\Delta(a) = 1$ when $a \geq 0$ and $\Delta(a) = 0$ otherwise gives the distribution function

$$\bar{H}_y = F(t) = N^{-1} \sum_{j \in \Omega} \Delta(t - y_j) \tag{2.2}$$

for each t. Rao (1994) has suggested a general class of estimators of H_y given by

$$\hat{H}_y = \sum_{i \in s} d_i(s) h(y_i) \tag{2.3}$$

where the basic weights $d_i(s)$ can depend both on s and $i (i \in s)$ and satisfy the design unbiasedness condition. The choice $h(y) = y$ in (2.3) gives Godambe's (1955) class of estimators of total. If $d_i(s) = \pi_i^{-1}$ then (2.3) reduces to Horvitz and Thompson (1952) estimator of population total. If $d_i(s) = w_i^*$ and $h(y_i) = I(y_i \leq t)$, then (2.3) reduces to the estimator $\hat{F}(t)$ suggested by Silva and Skinner (1995). Rao (1979) has suggested an estimator to estimate the variance of the estimator $\hat{H}_y$, i.e. $V(\hat{H}_y)$ as

$$\hat{V}(\hat{H}_y) = \sum \sum_{\substack{i<j \\ i,j \in s}} d_{ij}(s) w_i w_j (z_i - z_j)^2 \tag{2.4}$$

where $z_i = h(y_i)/w_i$ and weights $d_{ij}(s)$ can depend both on s and $(i, j) \in s$, and satisfy the unbiasedness condition. It is remarkable that (2.4) depends on the condition theory that $\hat{H}_y$ equals H_y where $h(y_i) \propto w_i$. The Yates and Grundy (1953) estimator of the variance of Horvitz and Thompson (1952) estimator is a special case of (2.4) with $w_i = \pi_i$ and $d_{ij}(s) = (\pi_i \pi_j - \pi_{ij})/(\pi_{ij} \pi_i \pi_j)$ for any fixed sample size, n, design. In fact, Rao (1994) has suggested an extension of the calibration approach proposed by Deville and Särndal (1992) to estimate any kind of central tendency parameter of the study variable using known central tendency parameter of the auxiliary character. Under the super population model, defined as

$$y_i = \beta x_i + \varepsilon_i \tag{2.5}$$

where β is an unknown constant, ε_i's are independently distributed random variables with means $E_m(\varepsilon_i) = 0$ and variance $V_m(\varepsilon_i) = \sigma_i^2$. If this model is tenable, then the well- known GREG predictor for estimating population total Y is

$$t_g = \sum_{i \in s} \frac{y_i}{\pi_i} + \hat{\beta} \left(X - \sum_{i \in s} \frac{x_i}{\pi_i} \right) \tag{2.6}$$

Särndal (1982) suggested two estimators to estimate the variance of the estimator t_g as

$$\hat{V}_1 = \sum\sum_{\substack{i<j \\ i,j\in s}} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2 \tag{2.7}$$

and

$$\hat{V}_2 = \sum\sum_{\substack{i<j \\ i,j\in s}} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{g_{is}e_i}{\pi_i} - \frac{g_{js}e_i}{\pi_i}\right)^2 \tag{2.8}$$

where $e_i = y_i - \hat{\beta}x_i$, $g_{is} = 1 + (X - \sum_{i\in s}\frac{x_i}{\pi_i})\frac{Q_i\pi_i x_i}{\sum_{i\in s} Q_i x_i^2}$ and Q_i is an assignable constant. One can easily see that the estimators of variance at (2.7) and (2.8) are of the form (2.4). The estimators of variance proposed by Rao and Vijayan (1977) and Särndal (1996) can also be shown as the special cases of the estimator given at (2.4).

The next section has been devoted to develop a new estimator of the variance of the estimator $\hat{H}_y$, using known second order moment of the estimator of auxiliary character $\hat{H}_x$.

3. Regression type estimator

We propose an estimator of variance of $\hat{H}_y$ as

$$\hat{V}_s(\hat{H}_y) = \sum\sum_{\substack{i<j \\ i,j\in s}} \psi_{ij}(s) w_i w_j (z_i - z_j)^2 \tag{3.1}$$

where $\psi_{ij}(s)$ are the modified weights and are as close as possible in an average sense for a given measure to the $d_{ij}(s)$ with respect to the calibration equation:

$$\sum\sum_{\substack{i<j \\ i,j\in s}} \psi_{ij}(s) w_i w_j (q_i - q_j)^2 = V(\hat{H}_x) \tag{3.2}$$

where $q_i = \frac{h(x_i)}{w_i}$ and $V(\hat{H}_x) = \sum\sum_{\substack{i<j \\ i,j\in\Omega}} d_{ij}(\Omega) w_i w_j (q_i - q_j)^2$ for $d_{ij}(\Omega) = \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i\pi_j}$ denote the known second order moment of the estimator, $\hat{H}_x = \sum_{i\in s} d_i(s)h(x_i)$, of the auxiliary parameter H_x. To compute the right hand side of (3.2), we need either information on every unit of auxiliary character in the population or $V(\hat{H}_x)$ known from a past survey or pilot survey. An example of a situation where information on every unit of the auxiliary character is known is establishment turnover recorded from census or administrative records. For example, Das and Tripathi (1978), Srivastava and Jhajj (1980, 1981), Garcia and Cebrian (1996), Shah and Patel (1996) and Mahajan and Singh (1996) have used known second order moments of the auxiliary variable at the estimation stage. Fuller (1970) has also given an idea to adjust the weights $d_{ij}(s)$ in the usual Yates and Grundy (1953) estimator of variance.

For simplicity we restrict ourselves to the two dimensional Chi-Square type distance D between two lower triangular $n \times n$ grids formed by the weights $\psi_{ij}(s)$ and $d_{ij}(s)$ for

$i, j = 1, 2, \ldots, n$ defined as

$$\sum_{\substack{i<j \\ i,j\in s}}\sum \frac{[\psi_{ij}(s) - d_{ij}(s)]^2}{d_{ij}(s)Q_{ij}} \tag{3.3}$$

In most of the situations $Q_{ij} = 1$ but other types of weights can also be used. We will show that the ratio type estimator is a special case for a particular choice of Q_{ij}. Minimization of (3.3) subject to (3.2) leads to the modified optimal weights given by

$$\psi_{ij}(s) = d_{ij}(s) + \frac{w_i w_j d_{ij}(s) Q_{ij}(q_i - q_j)^2}{\sum_{\substack{i<j \\ i,j\in s}}\sum w_i^2 w_j^2 d_{ij}(s) Q_{ij}(q_i - q_j)^4} \cdot \left[V(\hat{H}_x) - \sum_{\substack{i<j \\ i,j\in s}}\sum d_{ij}(s) w_i w_j (q_i - q_j)^2 \right]. \tag{3.4}$$

On substituting the value of $\psi_{ij}(s)$ from (3.4) in (3.1), we get a regression type estimator to estimate the variance of $\hat{H}_y$, given by

$$\hat{V}_s(\hat{H}_y) = \hat{V}(\hat{H}_y) + \hat{B}[V(\hat{H}_x) - \hat{V}(\hat{H}_x)] \tag{3.5}$$

where

$$\hat{B} = \frac{\sum_{\substack{i<j \\ i,j\in s}}\sum w_i^2 w_j^2 d_{ij}(s) Q_{ij}(q_i - q_j)^2 (z_i - z_j)^2}{\sum_{\substack{i<j \\ i,j\in s}}\sum w_i^2 w_j^2 d_{ij}(s) Q_{ij}(q_i - q_j)^4} \quad \text{and}$$

$$\hat{V}_s(\hat{H}_x) = \sum_{\substack{i<j \\ i,j\in s}}\sum d_{ij}(s) w_i w_j (q_i - q_j)^2$$

have their usual meanings. The leading term of the mean squared error of the proposed regression type estimator (3.5) is given by

$$MSE[\hat{V}_s(\hat{H}_y)] = V[\hat{V}(\hat{H}_y)] + B^2 V[\hat{V}(\hat{H}_x)] - 2B\,\mathrm{Cov}[\hat{V}(\hat{H}_y), \hat{V}(\hat{H}_x)] \tag{3.6}$$

where

$$B = \frac{\sum_{\substack{i<j \\ i,j\in\Omega}}\sum w_i^2 w_j^2 d_{ij}(\Omega) Q_{ij}(q_i - q_j)^2 (z_i - z_j)^2}{\sum_{\substack{i<j \\ i,j\in\Omega}}\sum w_i^2 w_j^2 d_{ij}(\Omega) Q_{ij}(q_i - q_j)^4} \tag{3.7}$$

and

$$\mathrm{Cov}[\hat{V}(\hat{H}_y), \hat{V}(\hat{H}_x)] = \sum_{\substack{i<j<k<l \\ i,j,k,l\in\Omega}}\sum\sum\sum\sum d_{ij}(\Omega) d_{kl}(\Omega)(\pi_{ijkl} - \pi_{ij}\pi_{kl})(q_i - q_j)^2 (z_i - z_j)^2 \tag{3.8}$$

where π_{ijkl} denotes the probability of including four units in the sample i.e. $\pi_{ijkl} = \text{pr}(i,j,k \& l \in s)$. Note that $\pi_{ijkl} = \pi_{ijk}$, when $i = l$ or $j = l$ or $k = l$ etc. Expression (3.6) shows that the proposed estimator $\hat{V}_s(\hat{H}_y)$ is better than the conventional estimator $\hat{V}(\hat{H}_y)$ if $B < 2\,\text{Cov}[\hat{V}(\hat{H}_y), \hat{V}(\hat{H}_x)]/V[\hat{V}(\hat{H}_x)]$ which holds in most of the practical situations. The proposed estimator is consistent because the ratio of modified weights to design weights [*i.e.* $\frac{\psi_{ij}(s)}{d_{ij}(s)}$] converges in design probability to unity. This condition is analogue of the condition given by Särndal *et al.* (1989) for one dimensional strategy. If we choose $Q_{ij} = (q_i - q_j)^{-2} w_i^{-1} w_j^{-1}$, then the estimator (3.5) reduces to ratio type estimator given by

$$\hat{V}_r(\hat{H}_y) = \hat{V}(\hat{H}_y) \left[\frac{V(\hat{H}_x)}{\hat{V}(\hat{H}_x)} \right]. \tag{3.9}$$

Remark 3.1. The proposed method provides an estimator for $V(\hat{H}_y)$ where $\hat{H}_y$ is a linear and homogeneous estimator for H_y. It is most important to note that the General Linear Regression Estimator (GREG) is not a homogeneous estimator. Thus it is evident that the method can be applied to estimate not only $V(\hat{H}_y)$ but also any variance of an H_y estimator (even if it is not homogeneous) as long as it is of the form

$$V = \sum_{i<j} d_{ij}(\Omega) w_i w_j \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2 \tag{3.10}$$

with $e_i = f(y_i, x_i)$ and therefore the formula $\hat{V}_s(\hat{Y}_{GREG})$ given at (5.5) can be deduced as a particular case. The author wish to thank a referee for prompting this observation.

4. Estimation of $MSE[\hat{V}_s(\hat{H}_y)]$

An estimator of the MSE of the proposed regression estimator (3.5) is suggested as

$$\widehat{MSE}[\hat{V}_s(\hat{H}_y)] = \hat{V}[\hat{V}(\hat{H}_y)] + \hat{B}^2 \hat{V}[\hat{V}(\hat{H}_x)] - 2\hat{B}\widehat{\text{Cov}}[\hat{V}(\hat{H}_y), \hat{V}(\hat{H}_x)] \tag{4.1}$$

where

$$\hat{B} = \frac{\sum\sum_{\substack{i<j \\ i,j \in s}} w_i^2 w_j^2 d_{ij}(s) Q_{ij} (q_i - q_j)^2 (z_i - z_j)^2}{\sum\sum_{\substack{i<j \\ i,j \in s}} w_i^2 w_j^2 d_{ij}(s) Q_{ij} (q_i - q_j)^4} \tag{4.2}$$

and

$$\widehat{\text{Cov}}[\hat{V}(\hat{H}_y), \hat{V}(\hat{H}_x)] = \sum_{\substack{i<j<k<l \\ i,j,k,l \in s}}^{n} \sum^{n} \sum^{n} \sum^{n} d_{ij}(s) d_{kl}(s) \frac{(\pi_{ijkl} - \pi_{ij}\pi_{kl})}{\pi_{ijkl}} (z_i - z_j)^2 (q_i - q_j)^2 \tag{4.3}$$

is an unbiased estimator of variance-covariance terms defined earlier.

5. Particular cases

We will like to show many types of estimation strategies available in the literature can be derived from the general estimator $\hat{V}_s(\hat{H}_y)$.

5.1 *Estimation of finite population variance*

Case 1. If $d_{ij}(s) = (\pi_i\pi_j - \pi_{ij})/(\pi_{ij}\pi_i\pi_j)$, $w_i = \pi_i$, $\pi_i = \pi_j = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ and $Q_{ij} = 1$ then, under Simple Random Sampling and Without Replacement (SRSWOR) design, (3.5) becomes

$$\hat{V}_S(\hat{Y}_{SRSWOR}) = \frac{N^2(1-f)}{n}[s_y^2 + \hat{b}(S_x^2 - s_x^2)] \tag{5.1}$$

where, $s_y^2 = \frac{1}{2n(n-1)}\sum_{i=1}^n\sum_{j=1}^n(y_i - y_j)^2$ is an unbiased estimator of $S_y^2 = (N-1)^{-1}\sum_{i=1}^N(y_i - \bar{Y})^2$ with $\bar{Y} = N^{-1}\sum_{i=1}^N y_i$, $s_x^2 = \frac{1}{2n(n-1)}\sum_{i=1}^n\sum_{j=1}^n(x_i - x_j)^2$ is an unbiased estimator of $S_x^2 = (N-1)^{-1}\sum_{i=1}^N(x_i - \bar{X})^2$ with $\bar{X} = N^{-1}\sum_{i=1}^N x_i$, $\hat{b} = \hat{\mu}_{22}/\hat{\mu}_{04}$, where

$$\hat{\mu}_{22} = \frac{N^4(1-f)}{n^4(n-1)}\sum_{i=1}^n\sum_{j=1}^n(y_i - y_j)^2(x_i - x_j)^2 \quad \text{and}$$

$$\hat{\mu}_{04} = \frac{N^4(1-f)}{n^4(n-1)}\sum_{i=1}^n\sum_{j=1}^n(x_i - x_j)^4$$

and $f = \frac{n}{N}$ is the finite population correction (fpc) factor in the SRSWOR design. The ratio $\hat{V}_S(\hat{Y}_{SRSWOR})/\{N^2(\frac{1-f}{n})\}$ is a regression type estimator of finite population variance S_y^2 as proposed by Isaki (1983).

Case 2. Under SRSWOR, the ratio $\hat{V}_r(\hat{H}_y)/\{N^2(\frac{1-f}{n})\}$ leads to the ratio type estimator, $s_I^2 = s_y^2(S_x^2/s_x^2)$, proposed by Isaki (1983).

5.2 *Estimation of variance of estimators of total*

Case 3. *Hansen and Hurwitz* (1943) *estimator*: If $d_{ij}(s) = (\pi_i\pi_j - \pi_{ij})/(\pi_{ij}\pi_i\pi_j)$, $w_i = \pi_i$, $\pi_i = nP_i$, $\pi_j = nP_j$, $\pi_{ij} = n(n-1)P_iP_j$, $Q_{ij} = 1$ then estimator (3.5) reduces to

$$\hat{V}_S(\hat{Y}_{PPSWR}) = \hat{V}(\hat{Y}_{HH}) + \hat{b}[V(\hat{X}_{HH} - \hat{V}(\hat{X}_{HH})] \tag{5.2}$$

where,

$$\hat{V}(\hat{Y}_{HH}) = \frac{1}{2n^2(n-1)}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{y_i}{P_i} - \frac{y_j}{P_j}\right)^2,$$

$$\hat{V}(\hat{X}_{HH}) = \frac{1}{2n^2(n-1)}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{x_i}{P_i} - \frac{x_j}{P_j}\right)^2,$$

$$V(\hat{X}_{HH}) = \frac{1}{n}\sum_{i=1}^N P_i\left(\frac{x_i}{P_i} - X\right)^2, \quad \hat{b} = \hat{\mu}_{22}/\hat{\mu}_{04} \quad \text{with}$$

$$\hat{\mu}_{22} = \frac{1}{n^4(n-1)}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{y_i}{P_i} - \frac{y_j}{P_j}\right)^2\left(\frac{x_i}{P_i} - \frac{x_j}{P_j}\right)^2 \quad \text{and}$$

$$\hat{\mu}_{04} = \frac{1}{n^4(n-1)}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{x_i}{P_i} - \frac{x_j}{P_j}\right)^4$$

have their usual meanings under probability proportional to size and with replacement (PPSWR) sampling.

Thus (5.2) represent a regression type estimator of variance of the well known estimator of population total proposed by Hansen and Hurwitz (1943). Then the estimator $\hat{V}_r(\hat{H}_y)$, reduces to the ratio type estimator, given by

$$\hat{V}_r(\hat{Y}_{PPSWR}) = \hat{V}(\hat{Y}_{HH}) \left[\frac{V(\hat{X}_{HH})}{\hat{V}(\hat{X}_{HH})}\right]. \tag{5.3}$$

Case 4. *Ratio estimator*: Under SRSWOR sampling design, the proposed strategy reduces to an estimator of the variance of the ratio estimator, given by

$$\hat{V}_S(\hat{Y}_{Ratio}) = \frac{N^2(1-f)}{n} \times \frac{1}{(n-1)} \sum_{i=1}^{n} e_i^2 \left(\frac{X}{\hat{X}}\right)^2 \left(\frac{S_x^2}{s_x^2}\right) \tag{5.4}$$

where $e_i = y_i - (\bar{y}/\bar{x})x_i$. The estimator (5.4) makes additional use of known variance S_x^2 of the auxiliary character than the class of estimators proposed by Deng and Wu (1987) for $g = 2$.

Case 5. *Regression estimator*: The proposed strategy reduces to the estimator of variance of the regression estimator of total under SRSWOR as

$$\hat{V}_S(\hat{Y}_{GREG}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2 + \hat{\psi}_1(X - \hat{X}) + \hat{\psi}_2(X - \hat{X})^2 + \hat{\psi}_3(S_x^2 - s_x^2) \tag{5.5}$$

where

$$\hat{\psi}_1 = \frac{(N-n)}{(\sum_{i=1}^{n} x_i^2)n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n}(e_i - e_j)(x_i e_i - x_j e_j),$$

$$\hat{\psi}_2 = \frac{(N-n)}{2N(n-1)(\sum_{i=1}^{n} x_i^2)^2} \sum_{j=1}^{n}\sum_{j=1}^{n}(x_i e_i - x_j e_j)^2$$

and

$$\hat{\psi}_3 = \frac{N^2(1-f)}{n\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^4} \left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{(x_i - x_j)(e_i - e_j) + \frac{(X - \hat{X})(x_i - x_j)^2}{\sum_{i=1}^{n} x_i^2}\right\}^2\right].$$

The estimator (5.5) has been recently suggested by Singh *et al.* (1998).

5.3 *Estimation of variance of estimators of distribution functions*

A ratio type estimator to estimate the variance of the post stratification estimator of distribution function defined by Silva and Skinner (1995) can easily be derived as the special case of the proposed strategy as

$$\hat{V}_r(a_i) = N^{-2} \sum_{\substack{i<j \\ i,j\in s}}\sum \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{a_i}{\pi_i} - \frac{a_j}{\pi_j}\right)^2 \tag{5.6}$$

$$\cdot \left[\frac{\sum\sum\limits_{\substack{i<j \\ i,j\in\Omega}} (\pi_i\pi_j - \pi_{ij}) \left(\frac{b_i}{\pi_i} - \frac{b_j}{\pi_j} \right)^2}{\sum\sum\limits_{\substack{i<j \\ i,j\in s}} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{b_i}{\pi_i} - \frac{b_j}{\pi_j} \right)^2} \right]$$

where a_i and b_i are the arguments corresponding to study variable y and auxiliary variable x as defined by Silva and Skinner (1995). Regression type estimator to estimate the variance of the post- stratified estimator proposed by Silva and Skinner (1995) can also be derived from the proposed general estimator $\hat{V}_S(\hat{H}_y)$.

6. Empirical study

In order to illustrate the performance of the proposed estimators of variance, we have considered the case of simple ratio estimator of total given by

$$\hat{Y}_{Ratio} = N\bar{y}(\bar{X}/\bar{x}). \tag{6.1}$$

Following Rao (1994), an estimator to estimate the variance of the estimator $\hat{Y}_{Ratio}$ can easily be derived as

$$\hat{V}_1(\hat{Y}_{Ratio}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2 \left[\frac{X}{\hat{x}} \right]^2 \tag{6.2}$$

whereas the proposed strategy reduces to the estimator of variance given by

$$\hat{V}_2(\hat{Y}_{Ratio}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^{n} e_i^2 \left[\frac{X}{\hat{x}} \right]^2 \left(\frac{S_x^2}{s_x^2} \right) \tag{6.3}$$

where $e_i = y_i - \left(\frac{\bar{y}}{\bar{x}}\right) x_i$.

Finite populations. For the purpose of numerical illustration, we have taken a population consisting of $N = 20$ units from Horvitz and Thompson (1952). The study variable, y, is the number of households in i-th block and known auxiliary character, x, is the eye estimated number of households in the i-th block. All possible samples of size $n = 5$ were selected by SRSWOR, which results in $\binom{N}{n} = 15504$ samples. For the k-th sample, the estimator $\hat{Y}_{Ratio}\,|_k$ at (6.1) was computed. Empirical mean squared error of this estimator was computed as

$$MSE(\hat{Y}_{Ratio}) = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} [\hat{Y}_{Ratio}\,|_k - Y]^2. \tag{6.4}$$

For the k-th sample, the ratio type estimators of variance $\hat{V}_h(\hat{Y}_{Ratio})\,|_k$, $h = 1, 2$, given by (6.2) and (6.3) respectively, for estimating the variance of the ratio estimator were also obtained. The bias in the h-th ratio type estimator of variance was computed as

$$B\{\hat{V}_h(\hat{Y}_{Ratio})\} = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} \hat{V}_h(\hat{Y}_{Ratio})\,|_k - MSE(\hat{Y}_{Ratio}) \tag{6.5}$$

Table 1. Comparison of $\hat{V}_2(\hat{Y}_{Ratio})$ with $\hat{V}_1(\hat{Y}_{Ratio})$.

n	$B[\hat{V}_1(\hat{Y}_{Ratio})]$	$B[\hat{V}_2(\hat{Y}_{Ratio})]$	RE	$CCI[\hat{V}_1(\hat{Y}_{Ratio})]$	$CCI[\hat{V}_2(\hat{Y}_{Ratio})]$
5	−211.33	217.01	166.57	0.93	0.95
6	−141.92	102.00	115.06	0.91	0.92
7	−99.34	58.60	109.23	0.90	0.90

Table 2. Comparison of $\hat{V}_2(\hat{Y}_{GREG})$ with $\hat{V}_1(\hat{Y}_{GREG})$.

n	$B[\hat{V}_1(\hat{Y}_{GREG})]$	$B[\hat{V}_2(\hat{Y}_{GREG})]$	RE	$CCI[\hat{V}_1(\hat{Y}_{GREG})]$	$CCI[\hat{V}_2(\hat{Y}_{GREG})]$
5	−328.49	−194.78	112.04	0.92	0.96
6	−223.92	−136.34	103.02	0.90	0.93
7	−157.88	−94.38	101.21	0.91	0.94

and mean squared error was computed as

$$MSE\{\hat{V}_h(\hat{Y}_{Ratio})\} = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} [\hat{V}_h(\hat{Y}_{Ratio}) \mid_k - MSE(\hat{Y}_{Ratio})]^2. \tag{6.6}$$

The percent relative efficiency of $\hat{V}_2(\hat{Y}_{Ratio})$ with respect to $\hat{V}_1(\hat{Y}_{Ratio})$ was calculated as

$$RE = MSE\{\hat{V}_1(\hat{Y}_{Ratio})\} \times 100/MSE\{\hat{V}_2(\hat{Y}_{Ratio})\}. \tag{6.7}$$

The coverage by 95% confidence intervals $CCI[\hat{V}_h(\hat{Y}_{Ratio})]$ for $h = 1, 2$ were calculated for h-th ratio type estimator of variance by counting the number of times the true population total, Y, falls between the limits defined as

$$\hat{Y}_{Ratio|_k} \mp t_{n-h-1}(a)\sqrt{\hat{V}_h(\hat{Y}_{Ratio}) \mid_k}. \tag{6.8}$$

These results were also obtained from all possible samples of size 6 and 7 and have been presented in Table 1. At the second stage of calibration, we are making use of an additional known parameter of the auxiliary character and hence loosing one degree of freedom.

A similar process was repeated for the regression estimator, $\hat{Y}_{GREG} \mid_k = \hat{Y} + \left(\sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2\right)(x - \hat{X})$, of total under a SRSWOR design. The biases, relative efficiency and CCI were obtained by using h-th estimator of variance of the regression estimator, $\hat{V}_h(\hat{Y}_{GREG}) \mid_k$ for $h = 1, 2$. The estimators $\hat{V}_h(\hat{Y}_{GREG})$ can easily be obtained from (6.2) and (6.3), respectively, by changing e_i as $e_i = y_i - \hat{\beta}_{x_i}$. The results obtained have been presented in Table 2. In addition, it was observed that for $n = 5$, 0.020% estimates of variance obtained from the estimator $\hat{V}_1(\hat{Y}_{GREG})$ and 0.022% estimates obtained from the estimator $\hat{V}_2(\hat{Y}_{GREG})$ were negative. Similar results were observed for more natural populations given by Sukhatme and Sukhatme (1970) and Cochran (1977).

Over all proposed estimators perform better than existing estimators. These results are obtained in FORTRAN-77 using PENTIUM-120 by following the guidelines of Bratley *et al.* (1983). From the above analysis, one can conclude that at the cost of

loosing one degree of freedom, the proposed test statistic is found to be more powerful than the existing test statistics.

In real life situations, the study variable and auxiliary variables may follow certain kind of distribution like normal, beta or gamma etc. In order to see the performance of the proposed strategies under such circumstances, we generated artificial populations and considered the problem of estimation of finite population variance through simulation.

Artificial populations. The size N of these populations is unknown. We generated a pair of n independent random numbers and y_i^* and x_i^* (say), $i = 1, 2, \ldots, n$, from a subroutine VNORM with PHI=0.6, seed(y)=8987878 and seed(x) = 2348789 following Bratley *et al.* (1983). For fixed $S_y^2 = 500$ and $S_x^2 = 200$, we generated transformed variables

$$y_i = \sqrt{S_y^2(1-\rho^2)}y_i^* + \rho S_y x_i^* \tag{6.9}$$

and

$$x_i = S_x x_i^* \tag{6.10}$$

for different values of the correlation coefficient ρ. Then

$$s_y^2 \mid_k = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$s_x^2 \mid_k = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad b \mid_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2(x_i - x_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^4}$$

were computed from the k-th sample, whereas $S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$ were also computed from the population. The proposed strategy yields an estimator of variance from the k-th sample as

$$\hat{v} \mid_k = s_y^2 \mid_k + \hat{b} \mid_k (S_x^2 - s_x^2 \mid_k). \tag{6.11}$$

The empirical variance of the usual estimator S_y^2 has been computed as

$$V(s_y^2) = \frac{1}{15000}\sum_{k=1}^{15000}[s_y^2 \mid_k - S_y^2]^2. \tag{6.12}$$

The empirical mean square error of the proposed estimator, $\hat{v} \mid_k$, was computed as

$$MSE(\hat{v}_k) = \frac{1}{15000}\sum_{k=1}^{15000}[\hat{v} \mid_k - S_y^2]^2 \tag{6.13}$$

and its empirical bias was computed as

$$B(\hat{v}_k) = \frac{1}{15000}\sum_{k=1}^{15000}\hat{v} \mid_k - S_y^2. \tag{6.14}$$

The percent relative efficiency (RE) of the proposed strategy with respect to usual estimator has been defined as

$$RE = V(s_y^2) \times 100/MSE(\hat{v}_k) \tag{6.15}$$

Table 3. Empirical results obtained from artificial normal populations generated by VNORM subroutine over 15000 iterations.

Sample size (n)	Correlation coefficient	RE(%)	Relative bias	CCI of proposed strategy	CCI of the usual estimator
50	0.5	92.43	0.0583	0.9929	0.9921
	0.6	101.15	0.0514	0.9819	0.9913
	0.8	154.23	0.0338	0.9719	0.9909
	0.9	269.43	0.0218	0.9712	0.9927
100	0.5	96.66	0.0338	0.9916	0.9915
	0.6	105.89	0.0286	0.9613	0.9928
	0.8	161.73	0.0161	0.9613	0.9812
	0.9	281.79	0.0078	0.9613	0.9822
500	0.5	99.35	0.0004	0.9832	0.9821
	0.6	108.42	0.0032	0.9502	0.9828
	0.8	163.45	0.0104	0.9502	0.9838
	0.9	279.77	0.0146	0.9502	0.9829
1000	0.5	99.38	0.0152	0.9651	0.9645
	0.6	107.93	0.0179	0.9501	0.9656
	0.8	160.60	0.0228	0.9501	0.9643
	0.9	269.55	0.0249	0.9501	0.9645
5000	0.5	99.92	0.0030	0.9534	0.9534
	0.6	107.57	0.0019	0.9500	0.9537
	0.8	150.88	0.0011	0.9500	0.9543
	0.9	226.85	0.0033	0.9500	0.9561

and the relative bias has been defined as

$$RB = B(\hat{v}_k)/\sqrt{MSE(\hat{v}_k)}. \tag{6.16}$$

The results obtained from the artificial normal populations are shown in the Table 3. It is observed that for fixed sample size, the percent relative efficiency (RE) is an increasing function of the positive correlation between the study variable and the auxiliary character. This also support the result given by Garcia and Cebrian (1996) for the case of normal populations. The 95% coverage by confidence intervals (CCI) for the usual estimators S_y^2 and proposed estimator $\hat{v}\mid_k$ were obtained by counting the number of times the true variance S_y^2 lies in the intervals given by $s_y^2\mid_k \mp F_{n-1,n-1}(\alpha)\sqrt{\hat{v}(s_y^2)}$ and $\hat{v}\mid_k \mp F_{n-1,n-1}(\alpha)\sqrt{\widehat{MSE}(\hat{v}_k)}$. Similar kind of results were observed from other distributions viz. beta, gamma etc.

Distribution function. A ratio type estimator to estimate the variance of the post stratification estimator of distribution function defined by Silva and Skinner (1995) can be derived as a special case of the proposed strategy as

$$\hat{V}(a_i, b_i) = \hat{V}_y(a_i)\left[\frac{\hat{V}_x(b_i)}{V_x(b_i)}\right] \tag{6.17}$$

Table 4. Comparison of estimators of variance of the distribution function for selecting good, better and best students from three schools having 53 students by using post-stratification mechanism over 20000 samples in each situation.

Classification	t_y	t_x	P_y	P_x	n	$CCI(\hat{V}(a_i))$	$CCI(\hat{V}(a_i, b_i))$
Best students	80	60	0.151	0.585	10	0.971	0.964
					12	0.973	0.951
					14	0.986	0.957
					16	0.975	0.971
					18	0.905	0.939
		65	0.151	0.377	10	0.980	0.901
					12	0.924	0.933
					14	0.949	0.957
					16	0.928	0.933
					18	0.730	0.756
Better students	75	60	0.321	0.585	10	0.890	0.927
					12	0.924	0.946
					14	0.949	0.927
					16	0.928	0.935
					18	0.730	0.769
		65	0.321	0.377	10	0.890	0.901
					12	0.924	0.933
					14	0.949	0.957
					16	0.928	0.933
					18	0.730	0.756
Good students	70	60	0.566	0.585	10	0.866	0.908
					12	0.901	0.921
					14	0.975	0.948
					16	0.755	0.951
					18	0.696	0.756
		65	0.566	0.377	10	0.876	0.921
					12	0.911	0.932
					14	0.975	0.955
					16	0.955	0.956
					18	0.924	0.932

*Good students includes both better and best students, but better students includes only best students.

where

$$\hat{V}_y(a_i) = N^{-2} \sum_{\substack{i<j \\ i,j \in s}} \sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{a_i}{\pi_i} - \frac{a_i}{\pi_j} \right)^2,$$

$$\hat{V}_x(b_i) = N^{-2} \sum_{\substack{i<j \\ i,j \in s}} \sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{b_i}{\pi_i} - \frac{b_j}{\pi_j} \right)^2 \quad \text{and}$$

$$V_x(b_i) = N^{-2} \sum_{\substack{i<j \\ i,j \in \Omega}} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{b_i}{\pi_i} - \frac{b_j}{\pi_j} \right)^2,$$

where a_i and b_i are the arguments corresponding to study variable y and auxiliary variable x is as defined by Silva and Skinner (1995). The estimator (6.17) has been compared with the estimator proposed by Silva and Skinner (1995) as $\hat{V}_y(a_i)$. In this case, we considered a purely hypothetical but very interesting example given in Gunst and Mason (1980) by taking three schools as three strata. The study variable Y was taken as "Grade 13 Average" and auxiliary character X was taken as "First year average". For simulation purposes, we first merged data from the three schools by putting a flag to each observation showing the school code 1, 2 or 3. Out of merged data of all $N = 53$ students, we selected different samples of size n units as shown in the Table 4 by SRSWOR scheme. Those selected students were post-stratified into three schools on the basis of the flag attached at the beginnings. We were interested to estimate the proportion of students getting "Grade 13 Average" more than 70, 75 and 80 marks. In other words, the value of t_y in the distribution functions $F_y(t_y)$ was set at 70, 75 and 80 showing the proportion of students falling in the category of students having good, better and best marks. From the "First Year Average" the proportion of students having marks more than $t_x = 60$ (or 65) marks was assumed to be known. The values of the arguments $a_i = I(y_i \geq t_y) - F_{yg(i)}(t_y)$ and $b_i = I(x_i \geq t_x) - F_{xg(i)}(t_x)$, where $F_{yg}(t_y) = N_{yg}^{-1} \sum_{i \in U_g} I(y_i \geq t_y)$ and $F_{xg}(t_x) = N_{xg}^{-1} \sum_{i \in U_g} I(x_i \geq t_x)$ denote the population distribution functions for Y and X respectively, were obtained. It was assumed that $V_x(b_i)$ is known. In this case if the 95% confidence intervals were obtained by counting the number of times the true proportion P_y lies in the confidence intervals given by, $\hat{F}_{ps}(t_y) \mp 1.96\sqrt{\hat{V}(a_i)}$ and $\hat{F}_{ps}(t_y) \mp 1.96\sqrt{\hat{V}(a_i, b_i)}$, respectively. The results have been presented in Table 4.

Acknowledgements

The author is thankful to the Associate Editor and two learned referees for asking very good and interesting questions on the original version of the manuscript and to bring it in the present form. Partial support of this research from the Natural Sciences and Engineering Research Council of Canada Grant A3111 is gratefully acknowledged.

References

Bratley, P., Fox, B. L. and Schrage, L. E. (1983). *A Guide to Simulation*, Springer, New York.

Chaudhuri, A. and Roy, D. (1997). Optimal variance estimation for generalized regression predictor, *J. Statist. Plann. Inference*, **60**, 139–151.

Cochran, W. G. (1977). *Sampling Techniques*, Wiley, New York.

Das, A. K. and Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance, *Sankhyā, C*, **40**, 139–148.

Deng, Lih-Yuan and Wu, C. F. J. (1987). Estimation of variance of the regression estimator, *J. Amer. Statist. Assoc.*, **82**, 568–576.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling, *J. Amer. Statist. Assoc.*, **87**, 376–382.

Fuller, W. A. (1970). Sampling with random stratum boundaries, *J. Roy. Statist. Soc. Ser. B*, **32**, 209–226.

Garcia, M. R. and Cebrian, A. A. (1996). Repeated substitution method: The ratio estimator for the population variance, *Metrika*, **43**, 101–105.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. Ser. B*, **17**, 269–278.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application. A data-oriented approach*, Marcel Dekker, New York.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations, *Ann. Math. Statist.*, **14**, 333–362.

Horvitz, D. G. and Thompson, D. J. (1952). A generalisation of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.*, **47**, 663–685.

Isaki, C. T. (1983). Variance estimation using auxiliary information, *J. Amer. Statist. Assoc.*, **78**(381), 117–123.

Mahajan, P. K. and Singh, S. (1996). An estimator of total in two stage sampling, *J. Statist. Res.*, **30**(1), 127–131.

Rao, J. N. K. (1979). On deriving mean square errors and their non- negative unbiased estimators in finite population sampling, *J. Indian Statist. Assoc.*, **17**, 125–136.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage, *Journal of Official Statistics*, **10**(2), 153–165.

Rao, J. N. K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size, *J. Amer. Statist. Assoc.*, **72**, 579–584.

Särndal, C. E. (1982). Implications of survey designs for generalized regression estimators of linear functions, *J. Statist. Plann. Inference*, **7**, 155–170.

Särndal, C. E. (1996). Efficient estimators with simple variance in unequal probability sampling, *J. Amer. Statist. Assoc.*, **91**, 1289–1300.

Särndal, C. E., Swensson, B. and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika*, **76**(3), 527–537.

Shah, D. N. and Patel, P. A. (1996). Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling, *Canad. J. Statist.*, **24**(3), 373-384.

Silva, P. L. D. Nascimento and Skinner, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification, *J. Official Statist.*, **11**(3), 277–294.

Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of the regression estimator: Higher level calibration approach, *Survey Methodology*, **24**, 41–50.

Srivastava, S. K. and Jhajj, H. S. (1980). A class of estimators using auxiliary information for estimating finite population variance, *Sankhyā, C*, **42**, 87–96.

Srivastava, S. K. and Jhajj, H. S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information, *Biometrika*, **68**, 341–343.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*, Iowa State University Press, Iowa.

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

Wu, C. F. J. (1982). Estimation of variance of the ratio estimator, *Biometrika*, **69**, 183–189.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size, *J. Roy. Statist. Soc. Ser. B*, **15**, 253–261.