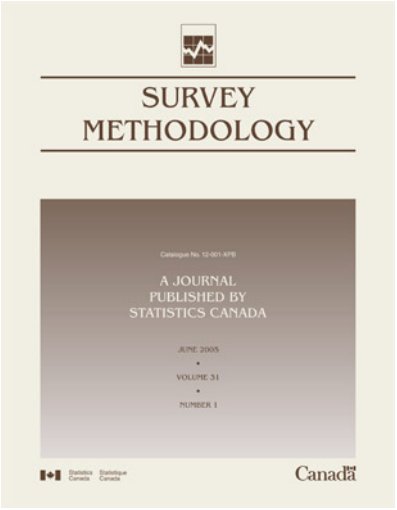
Catalogue no. 12-001-XIE

Survey Methodology

2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

2005

Published by authority of the Minister responsible for Statistics Canada

November 2005

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

4
Vol. 29, No. 1, pp. 99-104
Statistics Canada, Catalogue No. 12-001

Note on Calibration in Stratified and Double Sampling

D.S. Tracy, Sarjinder Singh and Raghunath Arnab [1]

Abstract

In the present investigation, new calibration equations making use of second order moments of the auxiliary character are introduced for estimating the population mean in stratified simple random sampling. Ways for estimating the variance of the proposed estimator are suggested, as well. The resultant new estimator can be more efficient than the combined regression estimator is in stratified sampling. The idea has been extended to double sampling in a stratified population and some simulation results studied.

Key Words: Calibration; Stratified Sampling; Double Sampling.

1. Introduction

Calibration estimation (Deville and Särndal 1992) has been much studied and practitioners have already offered many useful approaches (*e.g.*, Dupont 1995, Hidiroglou and Särndal 1998, Sitter and Wu 2002). Still more seems to remain to be done, as the use of this powerful technique expands further among practitioners.

This paper offers a modest extension of calibration estimation in the stratified and double sampling settings. We begin in this introduction by describing a new calibration estimator for the conventional stratified sample setting. Section 2 derives the variance of the proposed new estimator, followed by the derivation of a variance estimator. Section 3 extends these results to the important special case of double sampling. To explore the performance characteristics of the new estimator, some simulation results are presented in section 4 which concludes this brief note.

1.1 Standard Stratified Sampling Estimator

Suppose we have a population of N units that is first subdivided into L homogeneous subgroups called strata, such that the h^{th} stratum consists of N_h units, where $h = 1, 2, \ldots, L$ and $\sum_{h=1}^{L} N_h = N$. Suppose further that a sample of size n_h is drawn by Simple Random Sampling Without Replacement (SRSWOR) from the h^{th} population stratum such that $\sum_{h=1}^{L} n_h = n$, the required sample size. Finally, suppose the value of the i^{th} unit of the study variable selected from the h^{th} stratum is denoted by y_{hi}, where $i = 1, 2, \ldots, n_h$ and $W_h = N_h / N$ is the known proportion of population units falling in the h^{th} stratum.

In this standard set up (Cochran 1977), it can be shown that an unbiased estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1.1}$$

where $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ denotes the h^{th} stratum sample mean. Under SRSWOR sampling, the variance of the estimator $\bar{y}_{st}$ is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h} \right) S_{hy}^2 \tag{1.2}$$

where $S_{hy}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$ denotes the h^{th} stratum population variance, $\bar{Y}_h = N_h^{-1} \sum_{i=1}^{N_h} Y_{hi}$ denotes the h^{th} stratum population mean and $f_h = n_h / N_h$.

1.2 Proposed New Calibration Estimator

Let $X_{hi}, i = 1, 2, \ldots, N_h; h = 1, 2, \ldots, L$ denote the value of the i^{th} unit of the auxiliary variable in the h^{th} stratum about which information may be known at the unit level or at the stratum level. Consider a new alternative (calibration) estimator for stratified sampling of the form

$$\bar{y}_{st}(\text{new}) = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{1.3}$$

where the weights Ω_h are chosen such that the chi-square distance function

$$\sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{W_h Q_h} \tag{1.4}$$

where Q_h denotes suitable weights to form different forms of estimators such as combined ratio and combined regression type estimators, is minimized subject to the following two calibration constraints

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h \tag{1.5}$$

and

1. D.S. Tracy, Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, N9B 3P4, Canada; Sarjinder Singh, Department of Statistics, St. Cloud State University, 720 Fourth Avenue South, St. Cloud, MN 56301-4498, U.S.A. E-mail: sarjinder@yahoo.com; Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa. E-mail: arnab@pixie.udw.ac.za.

$$\sum_{h=1}^{L} \Omega_h s_{hx}^2 = \sum_{h=1}^{L} W_h S_{hx}^2, \qquad (1.6)$$

where $x_{hi}, i = 1, 2, \ldots, n_h; h = 1, 2, \ldots, L$ denotes the value of sampled i^{th} unit from the h^{th} stratum such that $\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi}$ denotes the h^{th} stratum sample mean estimator of the known h^{th} stratum population mean $\bar{X}_h = N_h^{-1} \sum_{i=1}^{N_h} X_{hi}$, and $s_{hx}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ denotes the h^{th} stratum sample variance estimator of the known h^{th} stratum population variance $S_{hx}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$ of the auxiliary variable.

Now it can be shown that minimization of (1.4) subject to (1.5) and (1.6) leads to new calibrated weights given by

$$\Omega_h = W_h + \left\{ W_h Q_h \bar{x}_h \left[\begin{array}{l} \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h s_{hx}^4 \\ - \sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h s_{hx}^2 \end{array} \right] \right\} \Bigg/ \left\{ \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2 \right)^2 \right\}$$

$$+ \left\{ W_h Q_h s_{hx}^2 \left[\begin{array}{l} \sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \\ - \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \end{array} \right] \right\} \Bigg/ \left\{ \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2 \right)^2 \right\}. \qquad (1.7)$$

On substituting (1.7) in (1.3), we get

$$\bar{y}_{st}(\text{new}) = \sum_{h=1}^{L} W_h [\bar{y}_h + \hat{\beta}_1 (\bar{X}_h - \bar{x}_h) + \hat{\beta}_2 (S_{hx}^2 - s_{hx}^2)] \qquad (1.8)$$

where

$$\hat{\beta}_1 = \left\{ \sum_{h=1}^{L} W_h Q_h \bar{x}_h \bar{y}_h \left[\begin{array}{l} \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h s_{hx}^4 \\ - \sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \end{array} \right] \right\} \Bigg/ \left\{ \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2 \right)^2 \right\}$$

and

$$\hat{\beta}_2 = \left\{ \left[\begin{array}{l} \sum_{h=1}^{L} W_h Q_h s_{hx}^2 \bar{y}_h \\ \left[\sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right. \\ \left. - \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right] \end{array} \right] \right\} \Bigg/ \left\{ \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2 \right)^2 \right\}.$$

Since the ratio $\Omega_h / W_h \to 1$ in probability, as the sample size in each stratum tends to infinity, the proposed estimator of the population mean is consistent.

Note that we are calibrating the estimates of the sample mean and the sample variance from each stratum, instead of each value of x_i, to the corresponding population parameters. Further note that if the population variance for each stratum is unknown, but the population means $\bar{X}_h$, $h = 1, 2, \ldots, L$ are known (or $\bar{X}$ is known), then it is advised to use only the single constraint (1.5).

2. Variance and Variance Estimation

While the new estimator $\bar{y}_{st}(\text{new})$ has been shown above to have acceptable asymptotic properties, what about the variance of the estimator and how does one go about estimating the variance? These questions are addressed in this section. We begin by looking (in subsection 2.1) at the variance of $\bar{y}_{st}(\text{new})$ and then go on to show how that variance can be estimated (in subsection 2.2).

2.1 Variance of New Estimator

The variance of the estimator $\bar{y}_{st}(\text{new})$ is given by

$$V(\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h} \right) S_{hy}^2 \left\{ 1 - \lambda_{h11}^2 - \frac{(\lambda_{h11}\lambda_{h03} - \lambda_{h12})^2}{\lambda_{h04} - 1 - \lambda_{h03}^2} \right\} \qquad (2.1)$$

where $\lambda_{hrs} = \mu_{hrs} / \mu_{h20}^{r/2} \mu_{02}^{s/2}$ and $\mu_{hrs} = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^r (X_{hi} - \bar{X}_h)^s$.

The expression (2.1) shows that the proposed estimator is always at least as efficient as the combined regression estimator in stratified sampling defined as

$$\bar{y}_{st}(c) = \sum_{h=1}^{L} W_h [\bar{y}_h + \hat{\beta} (\bar{X}_h - \bar{x}_h)] \qquad (2.2)$$

with variance

$$V(\bar{y}_{st}(c)) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h} \right) S_{hy}^2 \{1 - \lambda_{h11}^2\}. \qquad (2.3)$$

The variance $V(\bar{y}_{st}(\text{new}))$ can be written as

$$V(\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \in_{hi}^2 \qquad (2.4)$$

where

$$\in_{hi} = (Y_{hi} - \bar{Y}_h) - \beta_1 (X_{hi} - \bar{X}_h) - \beta_2 \{(X_{hi} - \bar{X}_h)^2 - \sigma_{hx}^2\} \quad (2.5)$$

with $\sigma_{hx}^2 = N_h^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$.

2.2 Estimation of the Variance

An estimator for estimating the variance $V(\bar{y}_{st}(\text{new}))$ is given by

$$\hat{V}_0(\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) \frac{1}{n_h - 3} \sum_{i=1}^{n_h} e_{hi}^2 \qquad (2.6)$$

where

$$e_{hi} = (y_{hi} - \bar{y}_h) - \hat{\beta}_1 (x_{hi} - \bar{x}_h) - \hat{\beta}_2 \{(x_{hi} - \bar{x}_h)^2 - s_{hx}^{*2}\} \quad (2.7)$$

with $s_{hx}^{*2} = n_h^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ being the maximum likelihood estimator of σ_{hx}^2.

We also consider a calibrated estimator of the variance defined as

$$\hat{V}_1(\bar{y}_{st}(\text{new}))_1 = \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) \frac{1}{n_h - 3} \sum_{i=1}^{n_h} e_{hi}^2 . \qquad (2.8)$$

The estimator proposed by Wu (1985) is a special case of this estimator.

3. Double Sampling

In this section we extend our stratified sampling results to the stratified double sampling case. In particular, suppose the population of N units consists of L strata such that the h^{th} stratum consists of N_h units and $\sum_{h=1}^{L} N_h = N$. From the h^{th} stratum of N_h units, draw a preliminary large sample of m_h units by SRSWOR sampling and measure the auxiliary character x_{hi} only. Select a sub-sample of n_h units from the given preliminary large sample of m_h units by SRSWOR sampling and measure both the study variable y_{hi} and auxiliary variable x_{hi}. Let $\bar{x}_h^* = m_h^{-1} \sum_{i=1}^{m_h} x_{hi}$ and $S_{hx}^{*2} = (m_h - 1)^{-1} \sum_{i=1}^{m_h} (x_{hi} - \bar{x}_h^*)^2$ denote the first phase sample mean and variance. Also let $\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi}, s_{hx}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ and $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}, s_{hy}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ denote the second phase sample mean and variances for the auxiliary and study characters, respectively. We are considering an estimator of the population mean in stratified double sampling as

$$\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h^* \bar{y}_h \qquad (3.1)$$

where W_h^* are the calibrated weights such that the chi-square distance

$$\sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h} \qquad (3.2)$$

where Q_h are predefined weights used to obtain to different types of estimators, is minimized subject to the constraints

$$\sum_{h=1}^{L} W_h^* \bar{x}_h = \sum_{h=1}^{L} W_h \bar{x}_h^* \qquad (3.3)$$

and

$$\sum_{h=1}^{L} W_h^* s_{hx}^2 = \sum_{h=1}^{L} W_h s_{hx}^{*2} \qquad (3.4)$$

where $W_h = N_h / N$ are known stratum weights. We then get the calibrated weights, for stratified double sampling, as

$$W_h^* = W_h + \left\{ \begin{bmatrix} W_h Q_h \bar{x}_h \\ \left[\sum_{h=1}^{L} W_h (\bar{x}_h^* - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h s_{hx}^4 \right] \\ - \sum_{h=1}^{L} W_h (s_{hx}^{*2} - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \end{bmatrix} \right\} \Bigg/ \left\{ \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right) \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^4 \right) - \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right)^2 \right\}$$

$$+ \left\{ \begin{bmatrix} W_h Q_h s_{hx}^2 \\ \left[\sum_{h=1}^{L} W_h (s_{hx}^{*2} - s_{hx}^2) \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right] \\ - \sum_{h=1}^{L} W_h (\bar{x}_h^* - \vec{x}_h) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \end{bmatrix} \right\} \Bigg/ \left\{ \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right) \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^4 \right) - \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right)^2 \right\}. \qquad (3.5)$$

Substitution of (3.5) in (3.1) leads to a new estimator of the population mean in stratified random sampling. Thus a calibrated estimator of the population mean in stratified double sampling is given by

$$\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}_1^* \left[\sum_{h=1}^{L} W_h (\bar{x}_h - \bar{x}_h^*) \right] + \hat{\beta}_2^* \left[\sum_{h=1}^{L} W_h (s_{hx}^2 - s_{hx}^{*2}) \right] \qquad (3.6)$$

where $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ have their usual meanings. It is to be noted that the estimator (3.6) makes the use of the estimated first phase variance of the auxiliary character while estimating the population mean. Thus the estimator (3.6) is different than the usual separate regression type estimator available in the literature.

Since the ratio $W_h^* / W_h \to 1$ in probability, as the second-phase sample size in each stratum tends to infinity, the proposed estimator is a consistent estimator of the

population mean. The conditional variance of the stratified double sampling estimator, $\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h^* \bar{y}_h$, is

$$V[\bar{y}_{st}(d) \mid W_h^*] = \sum_{h=1}^{L} W_h^{*2} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hy}^2 \qquad (3.7)$$

where $S_{hy}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$.

A conditionally unbiased estimator of $V[\bar{y}_{st}(d) \mid W_h^*]$ is

$$\hat{V}[\bar{y}_{st}(d) \mid W_h^*] = \sum_{h=1}^{L} W_h^{*2} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_{hy}^2 \qquad (3.8)$$

where $s_{hy}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$.

It may be noted that in the proposed strategy, there is no need to go for higher order calibration for estimating the variance, because the calibrated weights W_h^* already make use of the estimated first phase variance of the auxiliary character. The minimum variance of the stratified double sampling estimator $\bar{y}_{st}(d)$, to the first order of approximation, is given by

$$V(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h} \right) S_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h} \right) S_{hy}^2 \left\{ 1 - \lambda_{h11}^2 - \frac{(\lambda_{h11}\lambda_{h03} - \lambda_{h12})^2}{\lambda_{h04} - 1 - \lambda_{h03}^2} \right\} \right]. \qquad (3.9)$$

The variance of the stratified double sampling estimator $\bar{y}_{st}(d)$ can also be written as

$$V(\bar{y}_{st}(d)) \approx \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h} \right) S_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h} \right) \frac{1}{N_h} \sum_{i=1}^{N_h} \epsilon_{hi}^2 \right] \qquad (3.10)$$

where

$$\epsilon_{hi} = (Y_{hi} - \bar{Y}_h) - \beta_1 (X_{hi} - \bar{X}_h) - \beta_2 \{ (X_{hi} - \bar{X}_h)^2 - \sigma_{hx}^2 \}. \qquad (3.11)$$

An estimator of variance $V(\bar{y}_{st}(d))$ is given by

$$\hat{V}(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h} \right) s_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h} \right) \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}^2 \right] \qquad (3.12)$$

where $e_{hi} = (y_{hi} - \bar{y}_h) - \hat{\beta}_1 (x_{hi} - \bar{x}_h) - \hat{\beta}_2 \{ (x_{hi} - \bar{x}_h)^2 - s_{hx}^{*2} \}$ denotes the estimate of the residual term and $s_{hx}^{*2} = n_h^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ denotes the maximum likelihood estimator of σ_{hx}^2.

We suggest here a new estimator of the variance in stratified double sampling as

$$\hat{V}(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^{*2} \left[\left(\frac{1}{m_h} - \frac{1}{N_h} \right) s_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h} \right) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} e_{hi}^2 \right]. \qquad (3.13)$$

Clearly

$$\lim_{m_h \to N_h} \hat{V}(\bar{y}_{st}(d)) = \hat{V}(\bar{y}_{st}(\text{new})) \text{ because } \lim_{m_h \to N_h} W_h^* \to \Omega_h.$$

Note that in two-phase sampling, an estimate of population parameter of the auxiliary character based on first-phase sample information (large sample) will always be better than the corresponding estimate based on only second-phase sample information. One can refer to Hidiroglou and Särndal (1998) to see that calibration to an estimate of such an unknown quantity works well.

4. Early Simulation Results and Some Conclusions

To begin our study of the operating performance of the proposed estimator with respect to the usual combined regression estimator in stratified sampling, we performed a few simulation experiments. These are described below and then some overall observations are made to conclude the paper.

4.1 Simulation Results

The following procedure for doing the simulation experiment was adopted. We assumed that the population consists of three strata and within each stratum the population followed the distributions shown in Table 1.

In each stratum different transformations on x_{hi}^* and y_{hi}^* were made by examining all possible combinations of the correlation coefficients $\rho_h = 0.5, 0.7$ and 0.9 and sample sizes $n_h = 5, 10,$ and 15. The quantities $S_{1x} = 4.5$, $S_{2x} = 6.2$, $S_{3x} = 8.4$ and $S_{hy} = 4.8$ were fixed in each stratum.

We generated 50,000 populations each of size 75 units and having 25 units in each stratum. From each stratum, SRSWOR samples were drawn and an average of the empirical mean squared error of the combined regression estimator was computed as:

$$\text{MSE}(\bar{y}_{st}(c)) = \frac{1}{50{,}000} \sum_{j=1}^{50{,}000} \left[\left(\sum_{h=1}^{3} W_h (\bar{y}_h + \hat{\beta} (\bar{X}_h - \bar{x}_h) \right)_j - \bar{Y} \right]^2 \qquad (4.1)$$

Table 1
Characteristics of the Population

Population	Stratum 1	Stratum 2	Stratum 3
	$y_{1i} = 15 + \sqrt{(1-\rho_1^2)}\, y_{1i}^* + \rho_1 \frac{S_{1x}}{S_{1y}} x_{1i}^*$	$y_{2i} = 100 + \sqrt{(1-\rho_2^2)}\, y_{2i}^* + \rho_2 \frac{S_{2x}}{S_{2y}} x_{2i}^*$	$y_{3i} = 200 + \sqrt{(1-\rho_3^2)}\, y_{3i}^* + \rho_3 \frac{S_{3x}}{S_{3y}} x_{3i}^*$
	$y_{1i} = 50 + x_{1i}^*$	$y_{2i} = 150 + x_{2i}^*$	$y_{3i} = 100 + x_{3i}^*$
1	$f(z_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} z_{hi}^{*\alpha_h - 1} e^{-z_{hi}^*}, \alpha_h = 0.3;$ for $z_{hi}^* = x_{hi}^*;\ \alpha_h = 1.5$ for $z_{hi}^* = y_{hi}^*; h = 1, 2, 3$		
2	$f(y_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} y_{hi}^{*\alpha_h - 1} e^{-y_{hi}^*}, \alpha_h = 0.3;\ f(x_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_{hi}^{*2}}{2}}; h = 1, 2, 3$		
3	$f(x_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} x_{hi}^{*\alpha_h - 1} e^{-x_{hi}^*}, \alpha_h = 0.3;\ f(y_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_{hi}^{*2}}{2}}; h = 1, 2, 3$		
4	$f(z_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_{hi}^{*2}}{2}}$ for $z_{hi}^* = x_{hi}^*, z_{hi}^* = y_{hi}^*;\ h = 1, 2, 3$		

where $\bar{Y} = \frac{25\times 15 + 25\times 100 + 25\times 200}{75} = 100.5.$

Similarly the empirical mean squared error of the proposed estimator is given by

$$\mathrm{MSE}(\bar{y}_{st}(p)) = \frac{1}{50{,}000} \sum_{j=1}^{50{,}000} \left[\left(\sum_{h=1}^{3} W_h (\bar{y}_h + \hat{\beta}_1 (\bar{X}_h - \bar{x}_h) + \hat{\beta}_2 (S_{hx}^2 - s_{hx}^2)) \right)_j - \bar{Y} \right]^2. \quad (4.2)$$

The percent relative efficiency of the proposed estimator with respect to combined regression estimator is given by

$$\mathrm{RE} = \frac{\mathrm{MSE}(\bar{y}_{st}(c))}{\mathrm{MSE}(\bar{y}_{st}(p))} \times 100. \quad (4.3)$$

The results so obtained demonstrated a modest improvement over all combinations studied for all four populations. The range of improvements was about 4.46% to 13.08% with the median being 5.19%.

Several empirical studies were also carried out similar in structure to those presented above. In particular we were able to illustrate the extent to which our approach was more efficient than that considered by Singh, Horn and Yu (1998) in stratified sampling. Quite similar results were observed for the double sampling setting. Using the simulation program with $m_h = 20, h = 1, 2, 3,$ with the same four populations as described earlier, the median improvement was observed as 3.17%, 7.20%, 5.28%, and 3.12%, respectively.

4.2 Some Overall Observations

We are comfortable that our new calibration estimator will perform well in many settings. Our simulation results demonstrate this in several special cases. As with other calibration estimators, however, there has been an appeal at various points to asymptotic results. Such appeals raise concerns in small samples. For example in section 3 we stated that the ratio $W_h^* / W_h \to 1$ in probability. This allowed us to conclude that our new double sampling estimator was asymptotically unbiased. We recommend that such appeals be checked before our estimator is used in an application, possibly by employing simulation studies similar to those in this paper but for situations like those that are to be sampled in the practitioner's particular setting.

Acknowledgements

The authors' are thankful to the Associate Editor, the Assistant Editor and the two anonymous referees for their fruitful comments on the original version this manuscript. We are sorry to inform you that Professor Tracy passed away on 27/12/1998.

References

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Deville, J.C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-135.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

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

Sitter, R.R., and Wu, C. (2002). Efficient estimation of quadratic finite population functions. *Journal of the American Statistical Association*, 97, 535-543.

Wu, C.F.J. (1985). Variance estimation for combined ratio and combined regression estimators. *Journal of the Royal Statistical Society, B*, 47, 147-154.