Survey Methodology, June 2007
Vol. 33, No. 1, pp. 61-68
Statistics Canada, Catalogue No. 12-001

Mean - Adjusted bootstrap for two - Phase sampling

Hiroshi Saigo [1]

Abstract

Two-phase sampling is a useful design when the auxiliary variables are unavailable in advance. Variance estimation under this design, however, is complicated particularly when sampling fractions are high. This article addresses a simple bootstrap method for two-phase simple random sampling without replacement at each phase with high sampling fractions. It works for the estimation of distribution functions and quantiles since no rescaling is performed. The method can be extended to stratified two-phase sampling by independently repeating the proposed procedure in different strata. Variance estimation of some conventional estimators, such as the ratio and regression estimators, is studied for illustration. A simulation study is conducted to compare the proposed method with existing variance estimators for estimating distribution functions and quantiles.

Key Words: Double Sampling; Resampling; Variance estimation.

1. Introduction

Two-phase sampling or double sampling is a powerful tool for efficient estimation in surveys. Usually, a large-scale first phase sample is taken where auxiliary variables, correlated with the characteristics of interest and relatively easily obtained, are observed. Then, a small-scale sub-sample is chosen from the first phase sample to measure the characteristics of interest that are harder to obtain. At the estimation stage, the auxiliary variables at the first phase are employed to obtain an efficient estimator.

A closed-form sample variance formula for an estimator can be complicated or even unavailable under two-phase sampling. Consequently, resampling methods, such as the jackknife and bootstrap, are appealing for two-phase sampling. Rao and Sitter (1995) and Sitter (1997) studied the delete-1 jackknife approach to the ratio and regression estimators under two-phase sampling and found the method provides design-consistent variance estimation with desirable conditional properties given the auxiliary variables.

A weakness of the delete-1 jackknife is that it cannot handle quantile estimation. Moreover, it is not trivial how one can incorporate the finite population correction into the jackknife variance estimation under two-phase sampling (see Lee and Kim 2002 and Berger and Rao 2006). The bootstrap, on the other hand, eliminates these problems if properly formulated.

Several bootstrap methods for two-phase sampling have been proposed and studied. Schreuder, Li and Scott (1987), Biemer and Atkinson (1993) and Sitter (1997) considered similar bootstrap methods which provide consistent variance estimation when sampling fractions are negligible. Rao and Sitter (1997) proposed a rescaling bootstrap for high sampling fractions.

A disadvantage of the rescaling approach is that it cannot handle the estimation of distribution functions and quantiles. In this paper, we propose a mean-adjusted bootstrap for two-phase sampling that accommodates the estimation of distribution functions and quantiles. The method is simple and includes the existing ones for negligible sampling fractions as a special case. Recently, Kim, Navarro, and Fuller (2006) studied replication variance estimation without rescaling for two-phase sampling in a more generalized framework than that of this paper. Our method, however, is different in that it internally incorporates the finite population correction.

This paper is organized as follows. Section 2 presents the mean-adjusted bootstrap for two-phase sampling. Section 3 illustrates how the proposed method works for some conventional estimators. A simulation for estimating distribution functions and quantiles is conducted in Section 4. Section 5 discusses further applications of the mean-adjusted bootstrap. Concluding remarks are given in Section 6.

2. Mean - Adjusted bootstrap

For notational simplicity, we assume there is only one stratum. To extend our method to stratified sampling, repeat the same procedure independently in different strata to obtain a bootstrap sample (see Rao and Sitter 1997, pages 759-762).

Let P be the set of unit labels in a population of size N. Suppose a simple random sample without replacement (SRSWOR) of size n_{A+B} from P is taken and denote the sampled labels by $A+B$. The auxiliary variable (vector) x_i is observed for $i \in A+B$. Then take a second phase SRSWOR of size $n_A < n_{A+B}$ from $A+B$ and denote the sampled labels by A. The characteristic (vector) y_i is

1. Hiroshi Saigo, Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku Tokyo 169-8050, Japan.

measured for $i \in A$. Let $B = (A + B) - A$, $n_B = n_{A+B} - n_A$, $\mathbf{y}_A = \{y_i : i \in A\}$, $\mathbf{x}_A = \{x_i : i \in A\}$, and $\mathbf{x}_B = \{x_j : j \in B\}$. An approximately design-unbiased estimator of parameter θ is assumed to be written as $\hat{\theta} = t(\mathbf{y}_A, \mathbf{x}_A, \mathbf{x}_B)$.

Under the proposed method, a bootstrap sample is constructed as follows.

1. Regard A as an SRSWOR of size n_A from P. Choose n_A units from A by a bootstrap method suitable for an SRSWOR of size n_A from P. Denote the sampled labels by A^*.
2. Regard B as an SRSWOR of size n_B from $P - A$ conditional on A having been selected. Choose n_B units from B by a bootstrap method suitable for an SRSWOR of size n_B from $P - A$. Denote the sampled labels by B^*.
3. For $j \in B^*$, define the mean-adjustment as $\tilde{x}_j$, where

$$\tilde{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_{A^*})/(1 - f_A), \qquad (1)$$

with $\bar{x}_A = n_A^{-1}\sum_{i\in A} x_i$, $\bar{x}_{A^*} = n_A^{-1}\sum_{i\in A^*} x_i$, and $f_A = n_A / N$.

4. Let $\mathbf{y}_{A^*} = \{y_i : i \in A^*\}$, $\mathbf{x}_{A^*} = \{x_i : i \in A^*\}$, and $\tilde{\mathbf{x}}_{B^*} = \{\tilde{x}_j : j \in B^*\}$. The bootstrap analogue of $\hat{\theta}$ is then given by $\hat{\theta}^* = t(\mathbf{y}_{A^*}, \mathbf{x}_{A^*}, \tilde{\mathbf{x}}_{B^*})$.

For bootstrap methods for a finite population, see Shao and Tu (1995, Chapter 6). The Bernoulli Bootstrap (BBE) proposed by Funaoka, Saigo, Sitter and Toida (2006) is appropriate for our method because of a reason specified later. To obtain a bootstrap sample A^* in the BBE, we conduct random replacement for each i in A: keep (x_i, y_i) in the bootstrap sample with probability $p = \{1 - (1 - n_A^{-1})^{-1} \ (1 - f_A)\}^{1/2}$ or replace it with one randomly selected from A. For the case where $p \notin [0, 1]$, see Funaoka *et al.* (2006).

To estimate the variance of $\hat{\theta}$, repeat steps 1-4 a large number of times K and use

$$v_{\text{boot}}(\hat{\theta}) = K^{-1}\sum_{k=1}^{K}(\hat{\theta}^*_{(k)} - \hat{\theta}^*_{(\cdot)})^2, \qquad (2)$$

where $\hat{\theta}^*_{(k)}$ is the value of $\hat{\theta}^*$ in the kth bootstrap sample and $\hat{\theta}^*_{(\cdot)} = K^{-1}\sum_k \hat{\theta}^*_{(k)}$.

When f_A is negligible, the mean adjustment (1) is unnecessary. The above method then reduces for large n_A to that by Schreuder *et al.* (1987) and Sitter (1997).

The proposed bootstrap method is motivated by the following two observations. First, let sampling schemes I and II be $[P \to A + B,\ A + B \to A]$ and $[P \to A,\ P - A \to B]$, respectively, where $\to$ means "the right hand side is an SRSWOR from the left hand side." Then, I and II implement the identical design. In fact, the design probability assigned to a particular sample $\{\mathbf{i} = (i_1, i_2, ..., i_{n_A}) \in A,\ \mathbf{j} = (j_1, j_2, ..., j_{n_B}) \in B\}$ in I is $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_{A+B}} \times {}_{n_{A+B}} C_{n_A}]^{-1} = n_A!\, n_B!(N - n_{A+B})!/N!$ while it is $\Pr\{\mathbf{i} \in A,\ \mathbf{j} \in B\} = [{}_N C_{n_A} \times {}_{N-n_A} C_{n_B}]^{-1} = n_A! n_B!(N - n_{A+B})!/N!$ in II. Obviously, the sampling distribution of an estimator under repeated sampling depends on the sampling design. So, it is a matter of convenience to assume II is carried out even when I is employed.

Second, to motivate the mean adjustment (1), observe that the mean of x of the set $P - A$, or the conditional expectation of $\bar{x}_B$ under repeated sampling given A, is $\bar{X}_{P-A} = (\bar{X} - f_A\bar{x}_A)/(1 - f_A)$. The bootstrap analogue of $\bar{X}_{P-A}$ is given by $\bar{X}_{P-A^*} = (\bar{X} - f_{A^*}\bar{x}_{A^*})/(1 - f_{A^*})$. So, equation (1) amounts to $\tilde{x}_j = x_j - \bar{X}_{P-A} + \bar{X}_{P-A^*}$, a mean adjustment similar to that proposed by Rao and Shao (1992) in the context of hot deck imputation under the uniform response mechanism. This mean adjustment ensures appropriate correlations between x in A^* and x in B^* required for consistent variance estimation with high sampling fractions (see Rao and Sitter 1997, page 760). Note that the condition $n_A = n_{A^*}$ or $f_A = f_{A^*}$ is essential for cancelling out $\bar{X}$ in the mean adjustment. Therefore, the mean-adjusted bootstrap requires a bootstrap method for SRSWOR which retains the original sample size, such as the BBE.

It is shown in Appendix A that the proposed bootstrap method provides design-consistent variance estimation for the class of estimators studied by Rao and Sitter (1997). Since no rescaling is performed, the method also works for estimation of distribution functions. Under some regularity conditions for the population distribution function, it provides design-consistent variance estimates for quantiles.

3. Illustrations

3.1 Ratio estimator

To illustrate, let us first consider the ratio estimator $\bar{y}_r = r_A\bar{x}_{A+B}$, where $r_A = \bar{y}_A/\bar{x}_A$, $w_A = n_A/n_{A+B}$, and $\bar{x}_{A+B} = w_A\bar{x}_A + (1 - w_A)\bar{x}_B$. Let $\bar{y}_r^* = (\bar{y}_{A^*}/\bar{x}_{A^*})\{w_A\bar{x}_{A^*} + (1 - w_A)\bar{\tilde{x}}_{B^*}\}$, the bootstrap analogue of $\bar{y}_r$. Using the results in Appendix A with $h(\bar{y}_A, \bar{x}_A, \bar{x}_B) = (\bar{y}_A/\bar{x}_A)\{w_A\bar{x}_A + (1 - w_A)\bar{x}_B\}$, we may approximate variance of $\bar{y}_r^*$ under the proposed bootstrap method $V_*(\bar{y}_r^*)$ by

$$\begin{aligned} V_*(\bar{y}_r^*) \doteq\ & (\bar{x}_{A+B}/\bar{x}_A)^2 \frac{(1 - f_A)}{n_A}\hat{S}^2_{dA} \\ & + 2(\bar{x}_{A+B}/\bar{x}_A)\frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} \\ & + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)}\hat{S}^2_{xA} + \frac{(1 - w_A)}{(1 - f_A)}\hat{S}^2_{xB}\right], \end{aligned} \qquad (3)$$

where $\hat{S}_{dA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)^2$, $\hat{S}_{dxA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)(x_i - \bar{x}_A)$, $\hat{S}_{xA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^2$, and $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B)^2$. The right hand side of (3) can be described as a "bootstrap-linearization" variance estimator. We denote it by $v_{\mathrm{BL}}(\bar{y}_r)$. Note that $v_{\mathrm{BL}}(\bar{y}_r)$ is almost identical to the jackknife-linearization variance estimator by Rao and Sitter (1995),

$$v_{\mathrm{JL}}(\bar{y}_r) = (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 + 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dxA} + \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \hat{S}_{xA+B}^2, \quad (4)$$

where $\hat{S}_{xA+B}^2 = (n_{A+B} - 1)^{-1} \sum_{A+B} (x_i - \bar{x}_{A+B})^2$, which agrees with equation 4.8 of Demnati and Rao (2004), page 25. Since they are close to $v_{\mathrm{JL}}(\bar{y}_r)$, $V_*(\bar{y}_{lr})$, its Monte Carlo approximation $v_{\mathrm{boot}}(\bar{y}_r^*)$ and $v_{\mathrm{BL}}(\bar{y}_{lr})$ should perform well not only unconditionally but conditionally on $(\bar{x}_{A+B} / \bar{x}_A)$ as well. It is interesting to note that Taylor linearization in deriving $v_{\mathrm{BL}}(\bar{y}_r)$ is performed around the sample means, not the population means (see the comment made by Demnati and Rao 2004, page 21).

3.2 Regression estimator

We next consider the regression estimator. The estimator of the population mean is $\bar{y}_{lr} = \bar{y}_A + b_A(\bar{x}_{A+B} - \bar{x}_A) = \bar{y}_A + (1 - w_A) b_A(\bar{x}_B - \bar{x}_A)$, where $b_A = \hat{S}_{xyA} / \hat{S}_{xA}^2$ with $\hat{S}_{xyA} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)(y_i - \bar{y}_A)$. Let $\bar{y}_{lr}^* = \bar{y}_{A^*} + (1 - w_A) b_{A^*} (\bar{\bar{x}}_{B^*} - \bar{x}_{A^*})$. Using the results in Appendix A (see also Appendix B), we have

$$V_*(\bar{y}_{lr}^*) \doteq \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right] + z_A^2 \frac{(1 - f_A)}{n_A} m_{22} + 2 z_A \frac{(1 - f_A)}{n_A} m_{12} + 2 z_A \frac{(1 - f_{A+B})}{n_{A+B}} b_A m_{21} + 4 z_A^2 \frac{(1 - f_A)}{n_A} a_A b_A \bar{x}_A \hat{S}_{xA}^2, \quad (5)$$

where $z_A = n_A(\bar{x}_{A+B} - \bar{x}_A) / \{(n_A - 1) \hat{S}_{xA}^2\}$, $m_{pq} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^p e_i^q$, $e_i = y_i - \bar{y}_A - b_A(x_i - \bar{x}_A)$, and $a_A = \bar{y}_A - b_A \bar{x}_A$. We call the right hand side of (5) a bootstrap-linearization variance estimator of $\bar{y}_{lr}$ and denote it by $v_{\mathrm{BL}}(\bar{y}_{lr})$. The jackknife-linearization variance estimator for $\bar{y}_{lr}$ (Sitter 1997, page 781) is

$$v_{\mathrm{JL}}(\bar{y}_{lr}) = \frac{(1 - f_A)}{n_A} m_{02} + \frac{(1 - f_{A+B})}{n_{A+B}} b_A^2 \hat{S}_{xA+B}^2 + \frac{z_A^2}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A)^2 e_i^2}{(1 - c_i)^2} + \frac{2 z_A}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A) e_i^2}{(1 - c_i)} + \frac{2 z_A b_A}{n_A (n_{A+B} - 1)} \sum_{i \in A} \frac{(x_i - \bar{x}_A)(x_i - \bar{x}_{A+B}) e_i}{(1 - c_i)}, \quad (6)$$

where $c_i = n_A^{-1} + (x_i - \bar{x}_A)^2 / \{(n_A - 1) \hat{S}_{xA}^2\}$, the leverage values. From (5) and (6), $v_{\mathrm{boot}}(\bar{y}_{lr})$, $v_{\mathrm{BL}}(\bar{y}_{lr})$ and $v_{\mathrm{JL}}(\bar{y}_{lr})$ perform in a similar fashion conditionally provided that $f_{A+B} \doteq 0$, n_A is large enough for all c_i to be nearly zero and the last term on the right hand side of (5) is negligible.

3.3 Estimation of distribution functions

As an example, let us take the model-calibrated pseudo-empirical maximum likelihood estimator (ME) under two-phase sampling proposed by Wu and Luan (2003) defined by

$$\hat{F}_{\mathrm{ME}}(t) = \sum_{i \in A} \hat{p}_i I(y_i \le t), \quad (7)$$

where $\hat{p}_i$ maximizes the pseudo-likelihood function $\hat{l}(p) = \sum_A (N / n_A) \log p_i$ subject to (a) $\sum_A p_i = 1 (0 < p_i < 1)$; and (b) $\sum_A p_i g_i = n_{A+B}^{-1} \sum_{A+B} g_i$ where $g_i = g(x_i, t) = P(y \le t \mid x_i)$ under a certain working model. For example, we may assume $\log(g_i / (1 - g_i)) = x_i' \theta$ with variance function $V(g) = g(1 - g)$. Chen, Sitter and Wu (2002) showed a simple algorithm for computing $\hat{p}_i$. It can be shown (see Wu and Luan 2003) that under the two-phase sampling considered in this paper,

$$\hat{F}_{\mathrm{ME}}(t) = n_A^{-1} \sum_{i \in A} I(y_i \le t) + \left\{ n_{A+B}^{-1} \sum_{i \in A+B} g_i - n_A^{-1} \sum_{i \in A} g_i \right\} \beta + o_p(n_A^{-1/2}),$$

where $\beta = \sum_P (g_i - \bar{g}) I(y \le t) / \sum_P (g_i - \bar{g})^2$ with $\bar{g} = N^{-1} \sum_P g_i$. Note that this equation is not used in estimation, but it shows that the variance of $\hat{F}_{\mathrm{ME}}(t)$ can be estimated by the mean-adjusted bootstrap since $\hat{F}_{\mathrm{ME}}(t)$ is approximated by a regression-type estimator.

3.4 Quantile estimation

Quantile estimation can be obtained by directly inverting $\hat{F}(t)$ by $\hat{F}^{-1}(\alpha) = \inf\{t : \hat{F}(t) \ge \alpha\}$ for some $\alpha \in (0, 1)$. For example, if (7) is used, then a quantile estimate is given by $y_{(k)}$, where $y_{(k)}$ is the kth order statistic of y such that $\sum_{i=1}^{k-1} \hat{p}_{(i)} < \alpha$ and $\sum_{i=1}^{k} \hat{p}_{(i)} \ge \alpha$ (Chen and Wu 2002). Under some conditions specified in Chen and Wu (2002), a

Bahadur-type representation for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ can be established. Thus the mean-adjusted bootstrap variance estimator for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ is design-consistent. Note that no closed form variance estimator for $\hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ is available, but a consistent variance estimator based on Woodruff's interval estimation (Woodruff 1952) can be applied.

4. Simulation

4.1 Population and sampling

A simulation study was conducted to examine the mean-adjusted bootstrap variance estimator for the estimators in Section 3. We report here the results for estimating distribution functions and quantiles. The results for the ratio and regression estimators are available from the author upon request.

First, the auxiliary variable x for a finite population P of size $N = 2{,}000$ were generated as Gamma(1, 1). The characteristic variable y was then generated by $y_i = x_i + \sqrt{x_i}\, v_i$, where $v_i \sim N(0,\ 0.5^2)$. An SRSWOR $A + B$ of size $n_{A+B} = 800$ was taken from the population and then an SRSWOR A of size $n_A = 200$ was selected from $A + B$. The population was fixed throughout all simulation runs since we focus on design-based repeated-sampling properties.

4.2 Estimation of distribution functions

For the estimation of distribution functions, we took $\hat{F}_{\mathrm{ME}}(t)$ as an example. Other estimators, *e.g.*, Chambers and Dunstan (1986) and Rao, Kovar and Mantel (1990), can be handled similarly when an estimator is approximately design-unbiased. The working model for g in $\hat{F}_{\mathrm{ME}}(t)$ was assumed to be logit with binomial variance. The bootstrap variance estimator $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ was calculated with $K = 200$. The BBE was used in constructing a bootstrap sample. The total simulation runs were $M = 5{,}000$ while the true MSE of $\hat{F}_{\mathrm{ME}}(t)$ at a given t was estimated by 50,000 runs.

We compared $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ with three variance estimators: Wu and Luan's (2003) analytical estimator, the standard delete-1 jackknife and an *ad hoc* fpc-adjusted delete-1 jackknife. Wu and Luan's (2003) estimator is

$$v_a(\hat{F}_{\mathrm{ME}}(t)) = (n_{A+B}^{-1} - N^{-1})\hat{S}_I^2 + (n_A^{-1} - n_{A+B}^{-1})\hat{S}_D^2,$$

where the two $\hat{S}^2$ components are estimated respectively by

$$\hat{S}^2 = s^2 + \left[\frac{1}{n_{A+B}(n_{A+B}-1)}\sum_{j>i:i,\,j\in A+B} u_{ij} - \frac{1}{n_A(n_A-1)}\sum_{j>i:i,\,j\in A} u_{ij}\right]\hat{\beta}_F,$$

where $s^2 = \{n_A(n_A-1)\}^{-1}\sum_{i<j:i,\,j\in A} v_{ij}$, and $\hat{\beta}_F = \sum_{i<j:i,\,j\in A} u_{ij}v_{ij} / \sum_{i<j:i,\,j\in A} u_{ij}^2$ with u_{ij} and v_{ij} specified as follows: For $\hat{S}_I^2$, $v_{ij} = (I_i - I_j)^2$ and $u_{ij} = (\hat{g}_i - \hat{g}_j)^2$ with $I_i = I(y_i \le t)$ and $\hat{g}_i = \hat{g}(x_i, t)$ estimated in A; For $\hat{S}_D^2$, $v_{ij} = (\hat{D}_i - \hat{D}_j)^2$ and $u_{ij} = \hat{g}_i(1-\hat{g}_i) + \hat{g}_j(1-\hat{g}_j)$ with $\hat{D}_i = I_i - \hat{g}_i\hat{\beta}$, $\hat{\beta} = \sum_{i\in A} I_i(\hat{g}_i - \bar{\hat{g}}_A) / \sum_{i\in A}(\hat{g}_i - \bar{\hat{g}}_A)^2$ and $\bar{\hat{g}}_A = n_A^{-1}\sum_{i\in A}\hat{g}_i$.

The standard delete-1 jackknife formula is given by

$$v_J(\hat{\theta}) = \frac{(n_{A+B}-1)}{n_{A+B}}\sum_{j\in A+B}(\hat{\theta}_{(-j)} - \hat{\theta}_{(\cdot)})^2,$$

where $\hat{\theta} = \hat{F}_{\mathrm{ME}}(t)$, $\hat{\theta}_{(-j)}$ is the jth jackknife pseudo-estimate and $\hat{\theta}_{(\cdot)} = n_{A+B}^{-1}\sum_{j\in A+B}\hat{\theta}_{(-j)}$. Note that for $j \in A$, both y_j and x_j are deleted from the sample while for $j \in B$, only x_j is deleted (see Rao and Sitter 1995 and Sitter 1997). The *ad hoc* fpc-adjusted formula is $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t)) = (1 - f_{A+B})v_J(\hat{F}_{\mathrm{ME}}(t))$.

Table 1 shows the relative bias (%Bias) and the coefficient of variation (CV) of the four variance estimators for $\hat{F}_{\mathrm{ME}}(t_\alpha)$ (α = 0.10, 0.25, 0.50, 0.75, 0.90), where $F(t_\alpha) = \alpha$. Here, %Bias and CV were calculated as $\%\mathrm{Bias} = 100 \times (M^{-1}\sum_{m=1}^{M} v^{(m)} - \mathrm{MSE})/\mathrm{MSE}$ and $\mathrm{CV} = [M^{-1}\sum_{m=1}^{M}(v^{(m)} - \mathrm{MSE})^2]^{1/2}/\mathrm{MSE}$, respectively, where $v^{(m)}$ is a variance estimate in the mth simulation run. Table 1 demonstrates that $v_J(\hat{F}_{\mathrm{ME}}(t))$ is biased upward since the sampling fractions are not negligible, that $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$ is biased downward since the *ad hoc* adjustment factor $(1 - f_{A+B})$ is too small, and that both $v_a(\hat{F}_{\mathrm{ME}}(t))$ and $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ are approximately unbiased although the latter is slightly more unstable, as is typical for a resampling method.

Table 1 Variance estimation for the pseudo-empirical MLE $\hat{F}_{\mathrm{ME}}(t_\alpha)$

		α				
Estimator		**0.10**	**0.25**	**0.50**	**0.75**	**0.90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	0.27	-0.22	0.64	0.83	2.73
	CV	0.19	0.14	0.14	0.15	0.24
$v_a(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	-2.29	-2.03	-0.47	-1.95	-3.26
	CV	0.17	0.11	0.09	0.11	0.19
$v_J(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	14.24	17.29	22.98	23.80	24.97
	CV	0.24	0.21	0.25	0.27	0.36
$v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t_\alpha))$	%Bias	-31.45	-29.63	-26.21	-25.72	-25.02
	CV	0.33	0.30	0.27	0.27	0.30

Paralleling Royall and Cumberland (1981a, 1981b), we ordered the $M = 5{,}000$ simulated samples on the values of $\bar{x}_{A+B} - \bar{x}_A$, classified them into 20 consecutive groups of $G = 250$ in each of which the simulated conditional $\mathrm{MSE}(\mathrm{MSE}_c)$ and conditional mean of $v(E_c(v))$ were computed. Figure 1 shows MSE_c and $E_c(v)$ plotted against the group averages of $\bar{x}_{A+B} - \bar{x}_A$ for $t_{0.10}$ and $t_{0.90}$. It is seen that both $v_a(\hat{F}_{\mathrm{ME}}(t))$ and $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}(t))$ behave

similarly conditioned on $\bar{x}_{A+B} - \bar{x}_A$. The jackknife variance estimators, $v_J(\hat{F}_{\mathrm{ME}}(t))$ and $v_{Jfpc}(\hat{F}_{\mathrm{ME}}(t))$, though biased, track a trend in MSE_c.

4.3 Quantile estimation

By directly inverting $\hat{F}_{\mathrm{ME}}(t)$, we estimated the α quantile. To obtain $\hat{p}_i$ for $\hat{F}_{\mathrm{ME}}(t)$, we fixed t at $\hat{t}_\alpha$, where $\hat{t}_\alpha = \inf\{t: n_A^{-1}\sum_A I(y_i \le t) \ge \alpha\}$, an estimator using only $\{y_i : i \in A\}$. For variance estimation, $K = 1{,}000$ bootstrap samples were created. For comparison, we also computed the Woodruff variance estimator (Woodruff 1952 and Shao and Tu 1995, page 238),

$$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha)) = \left[\frac{\hat{F}_{\mathrm{ME}}^{-1}(\alpha + \zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}}) - \hat{F}_{\mathrm{ME}}^{-1}(\alpha - \zeta_{1-\kappa/2}\hat{\sigma}_{\hat{F}})}{2\zeta_{1-\kappa/2}}\right]^2,$$

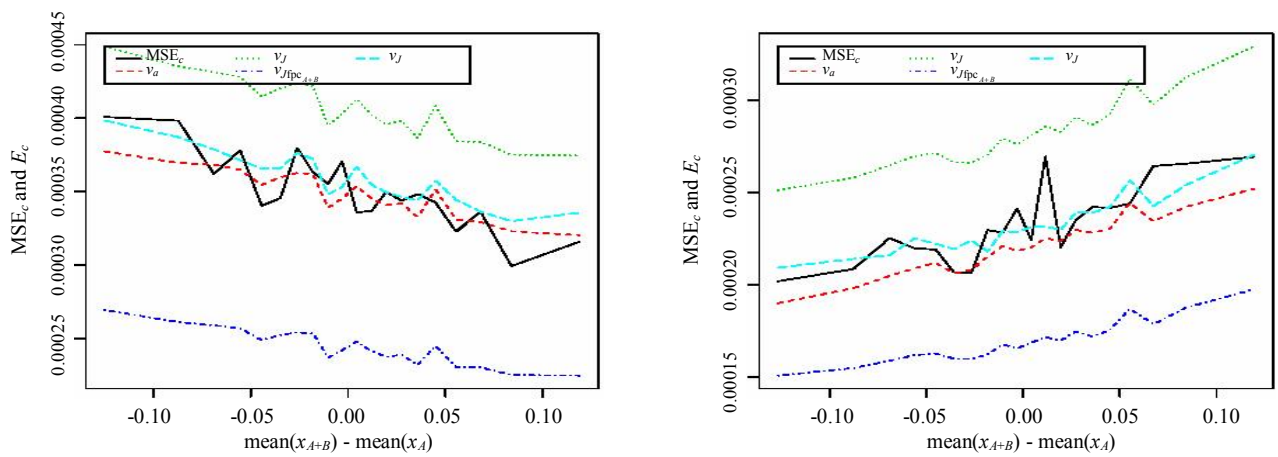
where $\hat{\sigma}^2_{\hat{F}} = v(\hat{F}_{\mathrm{ME}}(t))$ with $t = \hat{F}_{\mathrm{ME}}^{-1}(\alpha)$ and $\zeta_{1-\kappa/2}$ is the $(1-\kappa/2)$ quantile of $N(0, 1)$. We let $\kappa = 0.05$ although the best choice of κ is unknown. The performance measures, %Bias and CV, were calculated through $M = 5{,}000$ runs while the true MSE was estimated through 50,000 simulation runs.

Table 2 summarizes the results for quantile estimation. It demonstrates that the mean-adjusted bootstrap has an upward bias in estimating $V(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ while the bias in the Woodruff variance estimator is negligible.

Table 2 Variance estimation for quantiles

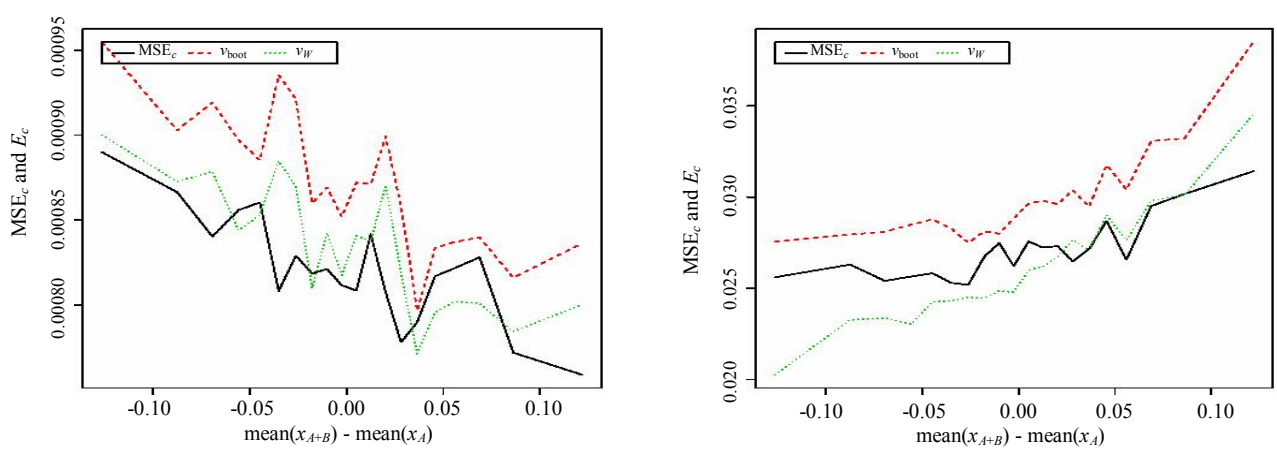
		α				
Estimator		**0.10**	**0.25**	**0.50**	**0.75**	**0.90**
$v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Bias	6.27	14.32	10.05	10.02	10.28
	CV	0.53	0.53	0.51	0.52	0.61
$v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$	%Bias	1.64	3.75	2.92	0.70	-3.67
	CV	0.50	0.45	0.45	0.46	0.52

Figure 2 shows conditional properties of $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ and $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ for $\alpha = 0.10, 0.90$. We see that both $v_{\mathrm{boot}}(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ and $v_W(\hat{F}_{\mathrm{ME}}^{-1}(\alpha))$ track MSE_c similarly although the former uniformly possesses an upward bias.



(a) $\hat{F}_{\mathrm{ME}}(t_{0.10})$ (b) $\hat{F}_{\mathrm{ME}}(t_{0.90})$

Figure 1 MSE_c and $E_c(v)$ for $\hat{F}_{\mathrm{ME}}(t_\alpha)$



(a) $\alpha = 0.10$ (b) $\alpha = 0.90$

Figure 2 MSE_c and $E_c(v)$ for quantile estimation

5. Further remarks

5.1 Stratified two-phase sampling

Suppose a population is to be stratified into H strata but no information for stratification is available. A possible solution for this situation is to first obtain an SRSWOR of size n' from the population, observe auxiliary variables including the ones for stratification, stratify the sample into H strata, and in each stratum take an SRSWOR of size n_h from n'_h units belonging to stratum h in the sample. See, for example, Cochran (1977, section 12.2) for details.

Let N_h be the size of stratum h in the population. Conditioned on $n'_h > 0$, the first-phase sampling in stratum h described above is equivalent to simple random sampling without replacement of size n'_h in stratum h independent across strata. Thus, given n'_h $(h = 1, ..., H)$, the mean-adjusted bootstrap can be applied independently in different strata to obtain a bootstrap sample. When N_h is unknown, as is usually the case for stratified two-phase sampling, an unbiased estimator $\hat{N}_h = N(n'_h/n')$ can be used in the mean-adjusted bootstrap. In this case, the sampling fraction n'/N is used commonly throughout all the strata.

Note, however, that the present discussion is legitimate for estimates conditioned on the first phase sample sizes. Variance due to the variable n'_h may be large. For unconditional variance estimation, see Kim *et al.* (2006).

5.2 Non-response

The above comment applies to imputed survey data under the uniform response mechanism. Let us suppose that a population is stratified into S_h $(h = 1, ..., H)$ where simple random sampling without replacement is undertaken independently. A sample is divided into imputation classes C_l $(l = 1, ..., L)$ in each of which the response rate is assume to be uniform and imputation is performed. An imputation class may cut across strata. We also assume which imputation class a sampled unit belongs to is correctly identified before imputation. Let us denote the numbers of sampled units and respondents in $S_h \cap C_l$ by n_{hl} and r_{hl}, respectively. Then, it is seen that given n_{hl} and r_{hl}, the corresponding design in $S_h \cap C_l$ is the same as the one discussed in this paper if we regard the n_{hl} units and r_{hl} respondents as $A + B$ and A, respectively. Therefore, the mean-adjusted bootstrap can be conducted independently in different $S_h \cap C_l$ $(h = 1, ..., H; l = 1, ..., L)$. The size of $S_h \cap C_l$, denoted by N_{hl}, can be estimated by $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Note that this is a bootstrap method conditioned on the number of respondents.

6. Conclusion

In this paper, we have proposed the mean-adjusted bootstrap for two-phase sampling. The method requires a simple mean adjustment and can handle the estimation of distribution functions and quantiles because it requires no rescaling. The Taylor series expansion shows that the method has desirable conditional properties for the ratio and regression estimators. A simulation study demonstrates that it also has similar conditional properties in estimating distribution functions and quantiles. An extension to stratified two-phase sampling is straightforward. Conditioned on the first phase sample sizes, the method can handle stratified two-phase sampling and imputation under the uniform response mechanism. We are currently invesigating an extension of the proposed method to more generalized multi-phase sampling designs.

Acknowledgements

This research was supported by a grant from the Japan Society of the Promotion of Science. The author would like to thank Professor Randy R. Sitter, the Editor, the Associate Editor and the two referees for their helpful comments and suggestions.

Appendix A

In this appendix, we show that the proposed bootstrap method provides consistent variance estimates for a class of estimators considered by Rao and Sitter (1997). We use the same setting as in Rao and Sitter (1997) with slightly different notation. For simplicity, we assume there exists only one stratum, but an extension to stratified two-phase sampling is straightforward.

Consider a class of estimators, $\theta = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$, of a population parameter $\theta = h(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, where $\bar{\boldsymbol{Y}}$ and $\bar{\boldsymbol{X}}$ are the population means of vectors $\boldsymbol{y}$ and $\boldsymbol{x}$, *i.e.*, $\bar{\boldsymbol{Y}} = N^{-1}\sum_{i\in P}\boldsymbol{y}_i$ and $\bar{\boldsymbol{X}} = N^{-1}\sum_{i\in P}\boldsymbol{x}_i$. Here, $\boldsymbol{x}$ is observed in the first phase sample $A + B$ whereas $\boldsymbol{y}$ is measured only in the second phase sample A. The sample means $(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A)$ and $\bar{\boldsymbol{x}}_B$ are calculated in A and B, respectively, *i.e.*, $\bar{\boldsymbol{y}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{y}_i$, $\bar{\boldsymbol{x}}_A = n_A^{-1}\sum_{i\in A}\boldsymbol{x}_i$, and $\bar{\boldsymbol{x}}_B = n_B^{-1}\sum_{i\in B}\boldsymbol{x}_i$.

By a Taylor expansion, we have

$$\hat{\theta} = \theta + \nabla h'(\Delta\bar{\boldsymbol{y}}'_A, \Delta\bar{\boldsymbol{x}}'_A, \Delta\bar{\boldsymbol{x}}'_B)' + o_p(n_A^{-1/2}),$$

where ∇h is the gradient vector of h evaluated at $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})$, $\Delta\bar{\boldsymbol{y}}_A = \bar{\boldsymbol{y}}_A - \bar{\boldsymbol{Y}}$, $\Delta\bar{\boldsymbol{x}}_A = \bar{\boldsymbol{x}}_A - \bar{\boldsymbol{X}}$, $\Delta\bar{\boldsymbol{x}}_B = \bar{\boldsymbol{x}}_B - \bar{\boldsymbol{X}}$, and $'$ means a transposed matrix (see equation 33.7 of Rao and Sitter 1997, page 757 and the required conditions therein). Then, the variance of $\hat{\theta} = h(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B)$ is approximated by

$$V(\hat{\theta}) \doteq \nabla h' \sum_{(\bar{\boldsymbol{y}}'_A, \bar{\boldsymbol{x}}'_A, \bar{\boldsymbol{x}}'_B)'} \nabla h,$$

where $\Sigma_{(\bar{y}'_A, \bar{x}'_A, \bar{x}'_B)'}$ is the variance-covariance matrix of $(\bar{y}'_A, \bar{x}'_A, \bar{x}'_B)'$ under repeated two-phase sampling. Because A and B are SRSWOR's of size n_A and n_B from the population P, respectively, we see that $\Sigma_{(\bar{y}'_A, \bar{x}'_A)'} = (1-f_A)S^2_{(y', x')'}/n_A$ and $\Sigma_{\bar{x}_B} = (1-f_B)S^2_x/n_B$, where $S^2_u = (N-1)^{-1}\Sigma_{i\in P}(u_i - \bar{U})(u_i - \bar{U})'$ is the population variance of $u = (y', x')'$ or x and $f_B = n_B/N$. For $\mathrm{Cov}(\bar{y}_A, \bar{x}_B)$, let E_A and $E_{B|A}$ be the expectation for selecting an SRSWOR A from P and choosing an SRSWOR B from $P-A$ given A, respectively. Note that $E_{B|A}(x_B) = (\bar{X} - f_A\bar{x}_A)/(1-f_A)$. So, we have

$$\begin{aligned}\mathrm{Cov}(\bar{y}_A, \bar{x}_B) &= E(\bar{y}_A\bar{x}'_B) - E(\bar{y}_A)E(\bar{x}'_B)\\ &= E_A(\bar{y}_A E_{B|A}(\bar{x}_B)) - \bar{Y}\bar{X}'\\ &= -S_{yx}/N,\end{aligned}$$

where $S_{yx} = (N-1)^{-1}\Sigma_{i\in P}(y-\bar{Y})(x-\bar{X})'$. Similarly, $\mathrm{Cov}(\bar{x}_A, \bar{x}_B) = -S^2_x/N$.

Now consider a Taylor expansion of $\hat{\theta}^* = h(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{\tilde{x}}_{B^*})$ with $\bar{\tilde{x}}_{B^*} = \bar{x}_{B^*} + f_A(\bar{x}_A - \bar{x}_{A^*})/(1-f_A)$, the bootstrap analogue of $\hat{\theta} = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$. Let E_* and V_* be the expectation and variance under the proposed bootstrap procedure, respectively. First, observe that $E_*(\bar{y}_{A^*}) = \bar{y}_A$, $E_*(\bar{x}_{A^*}) = \bar{x}_A$ and

$$\begin{aligned}E_*(\bar{\tilde{x}}_{B^*}) &= E_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{x}}_{B^*}))\\ &= E_{*A^*}(\bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*})/(1-f_A))\\ &= \bar{x}_B,\end{aligned}$$

where E_{*A^*} and $E_{*B^*|A^*}$ are respectively the expectation with respect to sampling A^* and the conditional expectation with respect to sampling B^* given A^* under the proposed bootstrap method. Then, $\hat{\theta}^* = h(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{\tilde{x}}_{B^*})$ is approximated by

$$\hat{\theta}^* = \hat{\theta} + \nabla h^{*'}(\Delta\bar{y}'_{A^*}, \Delta\bar{x}'_{A^*}, \Delta\bar{\tilde{x}}'_{B^*})' + o_p(n_A^{-1/2}),$$

where ∇h^* is the gradient of h evaluated at $(\bar{y}_A, \bar{x}_A, \bar{x}_B)$, $\Delta\bar{y}_{A^*} = \bar{y}_{A^*} - \bar{y}_A$, $\Delta\bar{x}_{A^*} = \bar{x}_{A^*} - \bar{x}_A$, and $\Delta\bar{\tilde{x}}_{B^*} = \bar{\tilde{x}}_{B^*} - \bar{x}_B$ (see equation 33.A.1 of Rao and Sitter 1997, page 767 and the required conditions therein). Therefore, $V_*(\hat{\theta}^*)$ is approximated by

$$V_*(\hat{\theta}^*) \doteq \nabla h^{*'}\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{\tilde{x}}'_{B^*})'}\nabla h^*,$$

where $\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{\tilde{x}}'_{B^*})'}$ is the variance-covariance matrix of $(\bar{y}_{A^*}, \bar{x}_{A^*}, \bar{\tilde{x}}_{B^*})'$ under the proposed bootstrap sampling.

Consistent variance estimation under the proposed method is proved by showing ∇h^* and $\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{\tilde{x}}'_{B^*})'}$ are consistent for ∇h and $\Sigma_{(\bar{y}'_A, \bar{x}'_A, \bar{x}'_B)'}$, respectively. Consistency of ∇h^* for ∇h follows from consistency of $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ for $(\bar{Y}, \bar{X}, \bar{X})$ and continuity of h.

Consistency of $\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{\tilde{x}}'_{B^*})'}$ can be shown as follows. First, since we use a bootstrap method suitable for simple random sampling without replacement in subsampling A^*, we have $\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*})'} = (1-f_A)\hat{S}^2_{(y', x')'A}/n_A$, where $\hat{S}^2_{uA} = (n_A-1)^{-1}\Sigma_{i\in A}(u_i - \bar{u}_A)(u_i - \bar{u}_A)'$ with $u = (y', x')'$. Second, because

1. $\Sigma_{\bar{\tilde{x}}_{B^*}} = E_{*A^*}(V_{*B^*|A^*}(\bar{\tilde{x}}_{B^*})) + V_{*A^*}(E_{*B^*|A^*}(\bar{\tilde{x}}_{B^*}))$, where V_{*A^*} and $V_{*B^*|A^*}$ are respectively the variance with respect to sampling A^* and the conditional variance with respect to sampling B^* given A^*,
2. $V_{*B^*|A^*}(\bar{\tilde{x}}_{B^*}) = (1-f_{B|A})\hat{S}^2_{xB}/n_B$, where $\hat{S}^2_{xB} = (n_B-1)^{-1}\Sigma_{i\in B}(x_i - \bar{x}_B)(x_i - \bar{x}_B)'$ and $f_{B|A} = n_B/(N-n_A)$, and
3. $E_{*B^*|A^*}(\bar{\tilde{x}}_{B^*}) = \bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*})/(1-f_A)$, we have $\Sigma^*_{\bar{\tilde{x}}_{B^*}} = (1-f_{B|A})\hat{S}^2_{xB}/n_B + f_A\hat{S}^2_{xA}/(N-n_A)$. Since both $\hat{S}^2_{xA}$ and $\hat{S}^2_{xB}$ are consistent for S^2_x, $\Sigma^*_{\bar{\tilde{x}}_{B^*}}$ is consistent for $\Sigma_{\bar{x}_B} = (1-f_B)S^2_x/n_B$. Finally, we compute $\mathrm{Cov}_*(\bar{y}_{A^*}, \bar{\tilde{x}}_{B^*})$ and $\mathrm{Cov}_*(\bar{x}_{A^*}, \bar{\tilde{x}}_{B^*})$. For the former, we have

$$\begin{aligned}\mathrm{Cov}_*(\bar{y}_{A^*}, \bar{\tilde{x}}_{B^*}) &= E_*(\bar{y}_{A^*}\bar{\tilde{x}}'_{B^*}) - E_*(\bar{y}_{A^*})E_*(\bar{\tilde{x}}_{B^*})'\\ &= E_{*A^*}(\bar{y}_{A^*}E_{*B^*|A^*}(\bar{\tilde{x}}_{B^*})) - \bar{y}_A\bar{x}'_B\\ &= E_{*A^*}(\bar{y}_{A^*}\{\bar{x}_B + f_A(\bar{x}_A - \bar{x}_{A^*})/(1-f_A)\}) - \bar{y}_A\bar{x}'_B\\ &= -\hat{S}_{yxA}/N,\end{aligned}$$

where $\hat{S}_{yxA} = (n_A-1)^{-1}\Sigma_{i\in A}(y_i - \bar{y}_A)(x_i - \bar{x}_A)'$. Similarly, $\mathrm{Cov}_*(\bar{x}_{A^*}, \bar{\tilde{x}}_{B^*}) = -\hat{S}^2_{xA}/N$. This completes the proof of consistency of $\Sigma^*_{(\bar{y}'_{A^*}, \bar{x}'_{A^*}, \bar{\tilde{x}}'_{B^*})'}$ for $\Sigma_{(\bar{y}'_A, \bar{x}'_A, \bar{\tilde{x}}'_B)'}$.

Appendix B

In this appendix, we derive $v_{\mathrm{BL}}(\bar{y}_{lr})$. Under the mean-adjusted bootstrap,

$$\begin{aligned}\bar{y}^*_{lr} &= \bar{y}_{A^*}\\ &+ (1-w_A)b_{A^*}\left\{-\frac{(\bar{x}_{A^*} - \bar{x}_A)}{(1-f_A)} + (\bar{x}_{B^*} - \bar{x}_B) + (\bar{x}_B - \bar{x}_A)\right\}.\end{aligned}$$

Define

$$\hat{\xi}^*_{pq} = n_A^{-1}\Sigma_{i\in A^*}x_i^p y_i^q,$$
$$\hat{\xi}^* = [\hat{\xi}^*_{10}, \hat{\xi}^*_{01}, \hat{\xi}^*_{11}, \hat{\xi}^*_{20}, \bar{x}_{B^*}]'$$

and

$$\xi = \left[\bar{x}_A, \bar{y}_A, n_A^{-1}\Sigma_{i\in A}x_i y_i, n_A^{-1}\Sigma_{i\in A}x_i^2, \bar{x}_B\right]' = E_*(\hat{\xi}^*).$$

Note that $b_{A^*} = (\hat{\xi}^*_{11} - \hat{\xi}^*_{10}\hat{\xi}^*_{01})/(\hat{\xi}^*_{20} - \hat{\xi}^{*2}_{10})$. Let $\bar{y}^*_{lr} = h(\hat{\xi}^*)$. This expression is slightly different from that in Appendix A, but we may exploit independent subsampling of A^* and B^*. Then, by Taylor linearization of $\bar{y}^*_{lr} = h(\hat{\xi}^*)$ around ξ,

we obtain $\bar{y}^*_{lr} \doteq \bar{y}_{lr} + \nabla h^{*'}(\hat{\xi}^* - \xi)$ and $V_*(\bar{y}^*_{lr}) \doteq \nabla h^{*'} \Sigma^*_{\hat{\xi}^*} \nabla h^*$, where

$$\nabla h^* = [-b_A(1-w_A)/(1-f_A) - z_A(\bar{y}_A - 2b_A\bar{x}_A), 1 - z_A\bar{x}_A, z_A, -z_A b_A, b_A(1-w_A)]'$$

and $\Sigma^*_{\hat{\xi}^*} = [v_{ij}]$ with

$$v_{11} = c_A \hat{S}^2_{xA},$$

$$v_{21} = c_A \hat{S}_{xyA},$$

$$v_{22} = c_A \hat{S}^2_{yA},$$

$$v_{31} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(x_i - \bar{x}_A),$$

$$v_{32} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(y_i - \bar{y}_A),$$

$$v_{33} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})^2,$$

$$v_{41} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i - \bar{x}_A),$$

$$v_{42} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(y_i - \bar{y}_A),$$

$$v_{43} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i y_i - \xi_{11}),$$

$$v_{44} = c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})^2,$$

$$v_{51} = v_{52} = v_{53} = v_{54} = 0,$$

$$v_{55} = \{n_B^{-1} - (N - n_A)^{-1}\} \hat{S}^2_{xB},$$

$v_{ij} = v_{ji}$, and $c_A = (1 - f_A)/n_A$. Rewriting the moments from the origin as the central moments, noting that $y_i - \bar{y}_A = b_A(x_i - \bar{x}_A) + e_i$ and using properties of e_i as the least-squares residuals, we obtain the right hand side of (5) after some algebra.

References

Berger, Y.G., and Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 68, 531-547.

Biemer, P.P., and Atkinson, D. (1993). Estimation of measurement bias using a model prediction approach. *Survey Methodology*, 19, 127-136.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.

Chen, J., and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition. New York: John Wiley & Sons, Inc.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.

Funaoka, F., Saigo, H., Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.

Kim, J.-K., Navarro, A. and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Lee, H., and Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K., and Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In *Survey Measurement and Process Quality*: *Wiley Series in Probability and Statistics*. (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York. 753-768.

Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Schreuder, H.T., Li, H.G. and Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag: New York.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Woodruff, R.S. (1952). Confidence intervals for median and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.