Statistica Sinica **23** (2013), 595-613
doi:http://dx.doi.org/10.5705/ss.2011.263

A JACKKNIFE VARIANCE ESTIMATOR FOR SELF-WEIGHTED TWO-STAGE SAMPLES

Emilio L. Escobar and Yves G. Berger

ITAM and University of Southampton

Abstract: Self-weighted two-stage sampling designs are popular in practice as they simplify field-work. It is common in practice to compute variance estimates only from the first sampling stage, neglecting the second stage. This omission may induce a bias in variance estimation; especially in situations where there is low variability between clusters or when sampling fractions are non-negligible. We propose a design-consistent jackknife variance estimator that takes account of all stages via deletion of clusters and observations within clusters. The proposed jackknife can be used for a wide class of point estimators. It does not need joint-inclusion probabilities and naturally includes finite population corrections. A simulation study shows that the proposed estimator can be more accurate than standard jackknifes (Rao, Wu, and Yue (1992)) for self-weighted two-stage sampling designs.

Key words and phrases: Linearisation, pseudovalues, Sen-Yates-Grundy form, smooth function of means, stratification.

1. Introduction

In survey sampling, accuracy of point estimates are assessed using variance estimates. Variance estimation becomes difficult when we have non-linear point estimators and complex sampling designs. This is a well-known problem which has been broadly covered in the survey sampling literature, e.g., Kish and Frankel (1974), Särndal, Swensson, and Wretman (1992), and Wolter (2007). Resampling techniques for variance estimation often overcome these difficulties. The Jackknife was first introduced by Quenouille (1956) for bias reduction and later by Tukey (1958) for variance estimation. This resampling technique has been widely studied, e.g. Krewski and Rao (1981), Kovar, Rao, and Wu (1988), Rao, Wu, and Yue (1992), and Shao and Tu (1995), among others.

Campbell (1980) proposed a totally different generalised jackknife variance estimator based on the analogy between linearisation and jackknife techniques. Berger and Skinner (2005) showed its design consistency for single stage designs under a set of regularity conditions. They also compared the empirical performance of Campbell's jackknife (in a single stage context) with standard single stage jackknifes such as in Tukey (1958), Kish and Frankel (1974), and Rao, Wu, and Yue (1992). Further, Berger and Rao (2006) extended Campbell's approach

for imputation. Berger (2007) proposed a modified Campbell's estimator which incorporates the Hájek (1964) approximation for the joint inclusion probabilities.

The regularity conditions in Berger and Skinner (2005) for the design-consistency of the Campbell estimator are too restrictive for two-stage sampling. For example, in two-stage simple random sampling the total number of sampled units would need to be fixed as population size tends to infinity for the Berger and Skinner (2005) regularity conditions to hold. In Section 3, we propose new and less restrictive regularity conditions that accommodate two-stage sampling. We also propose a Sen (1953) and Yates and Grundy (1953) version of Campbell's jackknife that overcomes the possibility of getting negative variance estimates. Further, the asymptotic design-consistency of these jackknife estimators is established under two-stage sampling.

In Section 4, we propose a jackknife variance estimator for self-weighted two-stage (stratified) without replacement sampling. These sampling designs are common in practice; examples include the Youth Risk Behavior Survey in the U.S.A., the Labour Force Survey for São Paulo in Brazil, and the Living Standards Survey for countries like South Africa, Ghana, and Côte d'Ivoire. We focus on self-weighted two-stage designs. However, there are different self-weighted designs that are widely used in practice. Some utilise three or more stages, and others use unequal probabilities at the final stage. Examples include the US National Health and Nutrition Examination Survey (NHANES) and the Australian and New Zealand Labour Force Surveys.

The proposed jackknife for self-weighted two-stage sampling involves deletion of both, clusters and observations. The proposed jackknife estimator does not have double sums and does not need joint inclusion probabilities. Further, we show that this novel estimator is asymptotically design-consistent. To ease computing efforts, a subsampling version is also proposed in Subsection 4.2 for its most computer intensive part that involves deleting observations.

In Section 5, Monte-Carlo simulations show that the proposed jackknife can be more accurate than customary jackknife estimators for more than one stage such as the Rao, Wu, and Yue (1992) stratified multi-stage delete-cluster jackknife.

2. The Class of Point Estimators

Let U denote a finite population of size N whose elements are grouped into N_I clusters of size M_i, $i = 1, \ldots, N_I$. Consider a without replacement sample s of elements drawn according to a self-weighted two-stage fixed sample size design. That is, n_I clusters are drawn using a without-replacement probability proportional to the size of the clusters, then a simple random sample without-replacement of m fixed elements is drawn within each sampled cluster. Therefore, the sample size is fixed and given by $n = n_I m$ elements grouped in n_I clusters.

Let $\pi_{Ii} > 0$ and π_{Iij} denote, respectively, the first and the second order inclusion probabilities for the clusters $i, j = 1, \ldots, N_I$; let $\pi_k > 0$ and $\pi_{k\ell}$ denote the inclusion probabilities for the elements $k, \ell = 1, \ldots, N$. For a self-weighted sampling design the clusters inclusion probabilities are $\pi_{Ii} = n_I M_i/N$, and thus $\pi_k = f$ where $f = n/N$.

Let y_{qk} denote the value of the survey variable q $(q = 1, \ldots, Q)$ for $k \in U$. Suppose we are interested in the population parameter

$$\theta = g(\mu_1, \ldots, \mu_Q),$$

a smooth and differentiable function of population means

$$\mu_q = \frac{1}{N} \sum_{k \in U} y_{qk}, \quad q = 1, \ldots, Q.$$

Further, assume θ is estimated by the substitution point estimator

$$\begin{aligned} \hat{\theta} &= g(\hat{\mu}_1, \ldots, \hat{\mu}_Q), \\ \hat{\mu}_q &= \sum_{k \in s} \tilde{w}_k y_{qk}, \quad q = 1, \ldots, Q, \end{aligned}$$

with $\hat{\mu}_q$ the Hájek (1971) mean estimator for μ_q with normalised sampling weights $\tilde{w}_k = w_k/\widehat{N}$, where $\widehat{N} = \sum_{k \in s} w_k$ and $w_k = 1/\pi_k$.

3. Generalised Jackknife Variance Estimators

The Campbell (1980) generalised jackknife variance estimator of $\hat{\theta}$ is (see Berger and Skinner (2005))

$$\widehat{\mathrm{var}}(\hat{\theta})_{HT} = \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\ \varepsilon_{(k)}\ \varepsilon_{(\ell)}, \tag{3.1}$$

with

$$\mathcal{D}_{k\ell} = \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}}, \qquad \varepsilon_{(k)} = (1 - \tilde{w}_k)(\hat{\theta} - \hat{\theta}_{(k)}), \tag{3.2}$$

where $\hat{\theta}_{(k)} = g(\hat{\mu}_{1(k)}, \ldots, \hat{\mu}_{Q(k)})$, $\hat{\mu}_{q(k)} = \sum_{\ell \in s-\{k\}} \tilde{w}_{\ell(k)} y_{q\ell}$, $\tilde{w}_{\ell(k)} = w_\ell (\sum_{\ell \in s-\{k\}} w_\ell)^{-1}$, and $s - \{k\}$ denoting s after deleting the k-th observation. Clearly the expression (3.1) may take negative values. To overcome this issue, we propose the alternative Sen (1953) and Yates and Grundy (1953) form,

$$\widehat{\mathrm{var}}(\hat{\theta})_{SYG} = \frac{-1}{2} \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\ (\varepsilon_{(k)} - \varepsilon_{(\ell)})^2, \tag{3.3}$$

which is always positive if the Sen-Yates-Grundy condition, $\mathcal{D}_{k\ell} < 0$, holds. Note that (3.3) is suitable for unequal-probability fixed sample size designs.

For single-stage sampling, Berger and Skinner (2005) showed the asymptotic design-consistency of (3.1) and also illustrated the better empirical performance of (3.1) in comparison with standard jackknifes such as in Tukey (1958), in Kish and Frankel (1974), and in Rao, Wu, and Yue (1992). Further, Berger and Rao (2006) extended (3.1) for imputation, and Berger (2007) proposed a modified version incorporating the Hájek (1964) approximation for the joint inclusion probabilities. Note that under uni-stage simple random sampling, both (3.1) and (3.3) reduce to the *standard jackknife* (e.g., Shao and Tu (1995, p.239); Wolter (2007)),

$$\widehat{\text{var}}(\hat{\theta})_{STD} = \left(1 - \frac{n}{N}\right) \frac{n-1}{n} \sum_{k \in s} (\hat{\theta}_{(k)} - \hat{\theta}_{(\cdot)})^2,$$

where $\hat{\theta}_{(\cdot)} = n^{-1} \sum_{k \in s} \hat{\theta}_{(k)}$.

3.1. Consistency of the generalised jackknifes for two-stage sampling

The consistency of $\widehat{\text{var}}(\hat{\theta})_{HT}$ and $\widehat{\text{var}}(\hat{\theta})_{SYG}$ is now to be set under new and less restrictive regularity conditions than those specified by Berger and Skinner (2005). These new conditions allow two-stage sampling.

We use the Isaki and Fuller (1982) asymptotic framework that considers a sequence of nested populations of size $N_{[t]}$ ($0 < N_{[t]} < N_{[t+1]}$), and a sequence of samples of size $n_{[t]}$ ($n_{[t]} < n_{[t+1]}, n_{[t]} < N_{[t]}$, for all t). To simplify notation, we drop the index t in what follows. Thus, $t \to \infty$ implies: $N \to \infty$, $n \to \infty$, and $n_I \to \infty$. We consider that $f = n/N$, $f_I = n_I/N_I$, and m are constants free of the limiting process.

For the vector of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_Q)^T$ and the vector of point estimators $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_Q)^T$, the multivariate Horvitz-Thompson and Sen-Yates-Grundy design variances and variance estimators of $\hat{\boldsymbol{\mu}}$ are approximated by (see Särndal, Swensson, and Wretman (1992, Secs. 5.5 and 5.7))

$$\begin{aligned}
\mathbf{var}(\hat{\boldsymbol{\mu}})_{HT} &\doteq \sum_{k \in U} \sum_{\ell \in U} \mathcal{D}_{k\ell}\ \pi_{k\ell}\ \boldsymbol{z}_k\ \boldsymbol{z}_\ell^T, \\
\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{HT} &\doteq \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\ \check{\boldsymbol{z}}_k\ \check{\boldsymbol{z}}_\ell^T, \\
\mathbf{var}(\hat{\boldsymbol{\mu}})_{SYG} &\doteq \frac{-1}{2} \sum_{k \in U} \sum_{\ell \in U} \mathcal{D}_{k\ell}\ \pi_{k\ell}\ \{\boldsymbol{z}_k - \boldsymbol{z}_\ell\}\{\boldsymbol{z}_k - \boldsymbol{z}_\ell\}^T, \\
\widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG} &\doteq \frac{-1}{2} \sum_{k \in s} \sum_{\ell \in s} \mathcal{D}_{k\ell}\ \{\check{\boldsymbol{z}}_k - \check{\boldsymbol{z}}_\ell\}\{\check{\boldsymbol{z}}_k - \check{\boldsymbol{z}}_\ell\}^T,
\end{aligned}$$

with $\boldsymbol{z}_k = N^{-1} w_k (\boldsymbol{y}_k - \boldsymbol{\mu})$, $\check{\boldsymbol{z}}_k = \tilde{w}_k (\boldsymbol{y}_k - \hat{\boldsymbol{\mu}})$, $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{Qk})^T$. Now, consider our regularity conditions.

C1. $\widehat{\text{var}}(\hat{\theta})_L/\text{var}(\hat{\theta})_L \to_p 1$, $\text{var}(\hat{\theta})_L \neq 0$, where

$$\begin{aligned} \text{var}(\hat{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T \, \mathbf{var}(\hat{\boldsymbol{\mu}})_{HT} \, \boldsymbol{\nabla}(\boldsymbol{\mu}), \\ \widehat{\text{var}}(\hat{\theta})_L &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T \, \widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{HT} \, \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}); \end{aligned} \tag{3.4}$$

for fixed sample-size designs,

$$\begin{aligned} \text{var}(\hat{\theta})_L &= \boldsymbol{\nabla}(\boldsymbol{\mu})^T \, \mathbf{var}(\hat{\boldsymbol{\mu}})_{SYG} \, \boldsymbol{\nabla}(\boldsymbol{\mu}), \\ \widehat{\text{var}}(\hat{\theta})_L &= \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T \, \widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG} \, \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}), \end{aligned} \tag{3.5}$$

where $\boldsymbol{\nabla}(\boldsymbol{x}) = (\partial g(\boldsymbol{\mu})/\partial\mu_1, \ldots, \partial g(\boldsymbol{\mu})/\partial\mu_Q)^T_{\boldsymbol{\mu}=\boldsymbol{x}}$ is the gradient of $g(\cdot)$ at $\boldsymbol{x} \in \Re^Q$, with $g(\cdot)$ continuous and differentiable at $\boldsymbol{\mu}$.

C2. $|1 - \tilde{w}_k| \geq \alpha > 0$, for all $k \in U$, α is a constant.

C3. $\liminf \{n \, \text{var}(\hat{\theta})_L\} > 0$.

C4. $n^{-1} \sum_{k \in s} \tilde{w}_k^\tau \, \|\boldsymbol{y}_k - \hat{\boldsymbol{\mu}}\|^\tau = \mathcal{O}_p(n^{-\tau})$ for all $\tau \geq 2$, where $\|\boldsymbol{A}\| = \text{tr}(\boldsymbol{A}^T\boldsymbol{A})^{1/2}$ denotes the Euclidean norm.

C5. $G_s = n^{-\beta} \sum \sum_{(k \neq \ell) \in s} (\mathcal{D}^-_{k\ell})^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}^-_{k\ell} = -\mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} < 0$ and 0 otherwise.

C6. $H_s = n^{-\beta} \sum \sum_{(k \neq \ell) \in s} (\mathcal{D}^+_{k\ell})^2 = \mathcal{O}_p(1)$, with $0 \leq \beta < 1$, where $\mathcal{D}^+_{k\ell} = \mathcal{D}_{k\ell}$ if $\mathcal{D}_{k\ell} \geq 0$ and 0 otherwise.

C7. $\boldsymbol{\nabla}(\boldsymbol{x})$ is Lipschitz continuous, $\|\boldsymbol{\nabla}(\boldsymbol{x}_1) - \boldsymbol{\nabla}(\boldsymbol{x}_2)\| \leq \lambda \, \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^\delta$, $\lambda > 0$ and $\delta > 0$ constants, $0 \leq \beta/2 < \delta$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the neighbourhood of $\boldsymbol{\mu}$.

C8. $\|\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})\| = \mathcal{O}_p(1)$.

Conditions C1 and C5 to C7 are similar, though different, to those in Berger and Skinner (2005); these conditions now allow two stage sampling. C1 sets the consistency of the linearisation variance estimator recalling the Robinson and Särndal (1983) approach (see Särndal, Swensson, and Wretman (1992, Secs. 5.5 and 5.7))). C2 to C4 are typical (e.g., Shao and Tu (1995, p.258)): C2 has that none of the normalised weights reach 1; C3 implies $\text{var}(\hat{\theta})_L$ decreases with rate n^{-1}; C4 is a Lyapunov-type condition for the existence of moments. C5 and C6 are mild conditions on the design, similar to ones in Isaki and Fuller (1982). C7 and C8 are standard smoothness requirements for the jackknife. Note that for two stage-sampling, $\beta < 1$ means that there are more observations than clusters in s.

Theorem 1. *For sampling designs of fixed size, if* C1 *to* C8 *hold, then the generalised jackknife variance estimator* $\widehat{var}(\hat{\theta})_{SYG}$ *at* (3.3) *is asymptotically design-consistent for the approximate linearised variance* $var(\hat{\theta})_L \neq 0$ *at* (3.5), $\widehat{var}(\hat{\theta})_{SYG}/var(\hat{\theta})_L \to_p 1$.

A proof of Theorem 1 is given in Appendix A.

Corollary 1. *If* C1 *to* C8 *hold, then* $\widehat{var}(\hat{\theta})_{HT}$ *at* (3.1) *is asymptotically design-consistent for the approximate linearised variance* $var(\hat{\theta})_L \neq 0$ *at* (3.4), $\widehat{var}(\hat{\theta})_{HT}/var(\hat{\theta})_L \to_p 1$.

Corollary 1 can be shown with the Berger and Skinner (2005) proof, taking into account the changes in the conditions C5 to C7. From Theorem 1, Corollary 1, and Slutsky's Theorem, when $\hat{\theta}$ is asymptotically normal,

$$\frac{\hat{\theta}-\theta}{(\widehat{\mathrm{var}}(\hat{\theta})_{SYG})^{1/2}} \to_d \mathrm{N}(0,1) \quad \text{and} \quad \frac{\hat{\theta}-\theta}{(\widehat{\mathrm{var}}(\hat{\theta})_{HT})^{1/2}} \to_d \mathrm{N}(0,1). \tag{3.6}$$

Thus, allowing valid confidence intervals for θ.

4. The Proposed Jackknife for Self-weighted Two-stage Sampling

For self-weighted two-stage sampling we have $\pi_{Ii} = n_I M_i/N$ and $\pi_k = n/N = f$. Now, by using the Hájek's approximation (Hájek (1964, eq. 5.27)) the clusters' joint inclusion probabilities π_{Iij} can be approximated by

$$\pi_{Iij} \doteq \pi_{Ii}\,\pi_{Ij}\,\{1 - d^{-1}(1-\pi_{Ii})(1-\pi_{Ij})\},$$

with $d = \sum_{Ii\in U}\pi_{Ii}(1-\pi_{Ii})$. This approximation was originally developed for $d \to \infty$, in our case $N_I \to \infty$, under the maximum entropy sampling design (see Hájek (1981, Them. 3.3, Chap. 3 and 6)), the Rejective Sampling design; a. k. a. the Conditional Poisson Sampling design. It requires that the utilised sampling design (of clusters) be of large entropy. An overview can be found in Berger and Tillé (2009). An account of different sampling designs, π_{Iij}'s approximations, and approximate variances under large-entropy designs can be found in Tillé (2006), Brewer and Donadio (2003), and Haziza, Mecatti, and Rao (2008). Recently, Berger (2011) gave sufficient conditions under which Hájek's results still hold for large entropy sampling designs that are not the maximum entropy one.

Low entropy sampling designs, such as the systematic probability proportional-to-size design, are not suitable for the above approximation. However, the randomized systematic sampling is suitable as it is of large entropy (e.g., Brewer and Gregoire (2009); Berger and Tillé (2009)).

Now, given the conditional inclusion probabilities $\pi_{k|Ii} = m/M_i$ and $\pi_{k\ell|Ii} = m(m-1)/M_i(M_i-1)$, the elements' joint inclusion probabilities $\pi_{k\ell}$ are

$$\pi_{k\ell} \doteq \begin{cases} \pi_{Ii}\,\pi_{k|Ii} = f & \text{if } (k=\ell)\in s_i, \\ \pi_{Ii}\,\pi_{k\ell|Ii} = f(m-1)/(M_i-1) & \text{if } (k\neq\ell)\in s_i, \\ \pi_{Iij}\,\pi_{k|Ii}\,\pi_{\ell|Ij} \doteq f^2\{1-d^{-1}(1-\pi_{Ii})(1-\pi_{Ij})\} & \text{if } k\in s_i, \ell\in s_j, i\neq j, \end{cases}$$

where s_i denotes the observations from the i-th cluster. Therefore, by substituting for $\mathcal{D}_{k\ell}$ in (3.2), we obtain

$$\mathcal{D}_{k\ell} \doteq \begin{cases} 1-f & \text{if } (k=\ell)\in s_i, \\ 1-\pi^*_{Ii} & \text{if } (k\neq\ell)\in s_i, \\ (1-\pi_{Ii})(1-\pi_{Ij})\{(1-\pi_{Ii})(1-\pi_{Ij})-d\}^{-1} & \text{if } k\in s_i, \ell\in s_j, i\neq j, \end{cases}$$

where

$$\pi^*_{Ii} = \pi_{Ii}\left(\frac{m}{m-1}\right)\left(\frac{M_i-1}{M_i}\right).$$

Thus, it can be shown (see Appendix C) that by substituting the values of $\mathcal{D}_{k\ell}$ into (3.3), it reduces to a jackknife variance estimator suitable for self-weighted two-stage sampling designs,

$$\widehat{\text{var}}(\hat\theta)_{prop} = v_{clu} \ + \ v_{obs}, \tag{4.1}$$

where

$$v_{clu} = \sum_{i\in s}(1-\pi^*_{Ii})\ \varsigma^2_{(Ii)} \ - \ \frac{1}{d}\left(\sum_{i\in s}(1-\pi_{Ii})\ \varsigma_{(Ii)}\right)^2, \tag{4.2}$$

$$v_{obs} = \sum_{k\in s}\phi_k\ \varepsilon^2_{(k)}, \tag{4.3}$$

with $\phi_k = \pi^*_{Ii}(M_i-m)(M_i-1)^{-1}$ for $k\in s_i$. Here the *delete cluster pseudo-values* $\varsigma_{(Ii)}$ are given by

$$\varsigma_{(Ii)} = \frac{n_I-1}{n_I}(\hat\theta-\hat\theta_{(Ii)}), \tag{4.4}$$

where $\hat\theta_{(Ii)} = g(\hat\mu_{1(Ii)},\ldots,\hat\mu_{Q(Ii)})$ with $\hat\mu_{q(Ii)} = \sum_{k\in s(Ii)}\tilde w_{k(Ii)}y_{qk}$, $\tilde w_{k(Ii)} = w_k(\sum_{k\in s(Ii)}w_k)^{-1}$, and $s(Ii) = s - s_i$ denoting the sample without observations from the i-th cluster; the *delete observation pseudo-values* $\varepsilon_{(k)}$, see Section 3, are given by

$$\varepsilon_{(k)} = \frac{n-1}{n}(\hat\theta-\hat\theta_{(k)}).$$

The proposed variance estimator (4.1) has two terms, one that deletes observations within clusters and another that deletes clusters. The term v_{clu} in (4.2) computes variability between clusters, and v_{obs} in (4.3) computes variability of observations within clusters. If d is unknown we can replace it by $\hat d = \sum_{i\in s}(1-\pi_{Ii})$. As $\phi_k \propto f(M_i-m)(m-1)^{-1}$. The term v_{obs} is zero if $m = M_i$, and it diminishes for small f. Conversely, it may become large if f is large, if the sampling fractions within clusters are small, or if the M_i vary.

To simplify notation, we consider (4.1) for non-stratified designs. This can be generalised by treating the strata separately. The number of strata has to be

bounded, and large sample regularity conditions must hold within each stratum. Therefore, the applicability of the proposed jackknife variance estimator excludes highly-stratified sampling designs with very few sampling units per stratum.

4.1. Consistency of the proposed jackknife

Let $\text{var}(\hat{\theta})_{HL}$ denote the Hájek approximation to the approximate linearised variance $\text{var}(\hat{\theta})_L$.

Theorem 2. *If* C1, C3, C4, C7, *and* C8 *hold, and if* $M_i \geq m \geq 2$*, then* $\widehat{var}(\hat{\theta})_{prop}$ *at* (4.1) *is asymptotically design-consistent for the Hájek approximate linearised variance* $var(\hat{\theta})_{HL} \doteq var(\hat{\theta})_L \neq 0$*,* $\widehat{var}(\hat{\theta})_{prop}/var(\hat{\theta})_L \to_p 1$*.*

A proof of Theorem 2 is given in Appendix B. Furthermore, (3.6) also holds for (4.1) when $\hat{\theta}$ is asymptotically normal.

4.2. A less computationally intensive version of the proposed jackknife

The delete-observation term v_{obs} at (4.3) can be laborious for large datasets. To ease computing, we propose to treat v_{obs} as a total that can be estimated from a subsample via the Horvitz and Thompson (1952) estimator. Hence we subsample $\tilde{n}$ elements from the sample s, $\tilde{s}$, and let $\tilde{\pi}_k$ be the first order inclusion probabilities of $\tilde{s}$. We estimate v_{obs} using the unbiased Horvitz-Thompson point estimator

$$\tilde{v}_{obs} = \sum_{k \in \tilde{s}} \frac{\phi_k \ \varepsilon^2_{(k)}}{\tilde{\pi}_k}. \tag{4.5}$$

Thus, a less computationally intensive estimator than (4.1) is given by

$$\widetilde{\text{var}}(\hat{\theta})_{prop} = v_{clu} \ + \ \tilde{v}_{obs}. \tag{4.6}$$

We recommend using $\tilde{\pi}_k = \tilde{n}\phi_k/\Phi$, where $\Phi = \sum_{k \in s} \phi_k$, implying

$$\tilde{v}^{\pi ps}_{obs} = \sum_{k=1}^{\tilde{n}} \frac{\phi_k \ \varepsilon^2_{(k)}}{\tilde{\pi}_k} = \frac{\Phi}{\tilde{n}} \sum_{k=1}^{\tilde{n}} \varepsilon^2_{(k)}.$$

Note that $\tilde{\pi}_k$ should be approximately proportional to $\phi_k \varepsilon^2_{(k)}$. Hence, this gives an efficient Horvitz-Thompson estimator.

In the context of two-phase sampling, Kim and Sitter (2003) also proposed a less computationally intensive approach. In further research, it would be good to explore the applicability of (4.6) for two-phase sampling designs.

4.3. Customary jackknife variance estimator

A customary jackknife variance estimator for sampling designs of more than one stage is the *stratified multi-stage delete cluster* jackknife estimator of Rao, Wu, and Yue (1992), originally purposed for functions of totals and for with-replacement sampling designs, thus for negligible sampling fractions. Their estimator is

$$\widehat{\mathrm{var}}(\hat{\theta})_{RWY} = \sum_{i \in s} \varsigma_{(Ii)}^2, \tag{4.7}$$

where $\varsigma_{(Ii)}$ is defined as (4.4). When the sampling fraction is large, this is usually adjusted by an overall clusters' finite population correction (FPC),

$$\widehat{\mathrm{var}}(\hat{\theta})_{RWY}^{FPC} = \sum_{i \in s} \left(1 - \frac{n_I}{N_I}\right) \varsigma_{(Ii)}^2. \tag{4.8}$$

Comparing (4.1) with (4.8), we note that $\widehat{\mathrm{var}}(\hat{\theta})_{prop}$ adds the term v_{obs} which computes variability of observations within clusters, and $\widehat{\mathrm{var}}(\hat{\theta})_{prop}$ uses a different FPC $(1 - \pi_{Ii}^*)$ for each cluster i, whereas $\widehat{\mathrm{var}}(\hat{\theta})_{RWY}^{FPC}$ uses the fixed FPC $(1 - n_I/N_I)$.

5. Simulation Study

We illustrate with two datasets: the Labour Force Population from Valliant, Dorfman, and Royall (2000, Appendix B.5) and the MU284 Swedish Municipalities Population from Särndal, Swensson, and Wretman (1992, Appendix B). For each, we duplicated 3 times the number of clusters and 3 times the number of observations within each cluster. We therefore used two population frames of $N = 4,302$ and $2,556$ observations, grouped into $N_I = 345$ and 150 clusters, respectively. The minimum/maximum cluster sizes were: 6/39 and 15/27, respectively. We used two variables from each population frame: weekly wages (y_1) and number of hours worked per week (y_2) from the first population frame, the number of Social-Democratic seats in municipal council (y_3) and the number of Conservative seats in municipal council (y_4) from the second. The homogeneity measures $\mho(\cdot)$ defined in Särndal, Swensson, and Wretman (1992, Secs. 3.4.3 and 4.2.2), for each of the variables of interest were: $\mho(y_1) = 0.2965$, $\mho(y_2) = 0.1951$, $\mho(y_3) = 0.3181$, and $\mho(y_4) = 0.4958$. The parameters of interest were the ratios $R_{12} = \mu_1/\mu_2 = 7.697$ and $R_{34} = \mu_3/\mu_4 = 2.439$, estimated by $\hat{R}_{12} = \hat{\mu}_1/\hat{\mu}_2$ and $\hat{R}_{34} = \hat{\mu}_3/\hat{\mu}_4$, where $\hat{\mu}_1, \ldots, \hat{\mu}_4$ are Horvitz-Thompson point estimators.

Clusters were selected using the Brewer (1975) unequal-probability sampling design with clusters' inclusion probabilities proportional to the cluster size; then, a simple random without replacement sample of individuals was selected within clusters using sample size $m = 2,\ 4$, and 6. For the labour force population frame,

it is important to note that 20.34% of the clusters are of the minimum cluster size, with $m = 6$ we are selecting all the elements within many clusters. For the estimator (4.6), we used the Brewer (1975) unequal probability design with subsampling rate 0.25 and with subsampling inclusion probabilities proportional to ϕ_k, as defined in Subsection 4.2.

For each simulation and each simulation example, $N_{Sim1} = 100,000$ and $N_{Sim2} = 1,000,000$ samples were selected to compute the empirical relative bias RB $= \mathrm{B}(\widehat{\mathrm{var}}(\hat{R}_{ab}))/\mathrm{var}(\hat{R}_{ab})$, where $\mathrm{B}(\widehat{\mathrm{var}}(\hat{R}_{ab})) = \mathrm{E}(\widehat{\mathrm{var}}(\hat{R}_{ab})) - \mathrm{var}(\hat{R}_{ab})$, the empirical relative root mean square error defined by RRMSE $= \{\mathrm{MSE}(\widehat{\mathrm{var}}(\hat{R}_{ab}))\}^{1/2}/\mathrm{var}(\hat{R}_{ab})$, and the coverage at a 95% confidence level. The $\mathrm{var}(\hat{R}_{ab})$ is the empirical variance computed from the N_{Sim1} (and N_{Sim2}) observed values of $\hat{R}_{ab}$ (ab =12 and 34). These quantities were computed for the estimators (4.1), (4.6), (4.7) and (4.8).

5.1. Example 1: point estimator $\hat{R}_{12}$

Results for this example are summarised in Table 1 which illustrates, in terms of RB, that the customary estimators (4.7) and (4.8), respectively, over-estimate and under-estimate the variance for increasing values of f (f_I). In general, this effect is more pronounced for small second-stage sampling sizes $m = 2, 4$, and less so with $m = 6$ (census within several clusters) for the FPC-adjusted customary estimator (4.8). On the other hand, the RB for the proposed variance estimator (4.1) and its subsampling version (4.6) remains close to zero as the sampling fractions increases, regardless of the second-stage sample sizes. The reason for this is that the proposed estimators correctly incorporate the finite population corrections at both stages. Note that there is a particular FPC for each cluster at (4.1) and (4.6).

In terms of RRMSE, it can be seen in Table 1 that the proposed estimator (4.1) had always the smallest RRMSE followed by the FPC-adjusted Rao, Wu, and Yue (1992) from (4.8), and by the subsampling version of the proposed estimator (4.6). It can also be seen in Table 1 that the original Rao, Wu, and Yue (1992) (4.7) had the correct coverage for small sampling fractions, although this variance estimator had the worst performance in terms of RB and RRMSE. Hence, discarding (4.7), the proposed estimator (4.1) has presumably the best coverage for increasing sampling fractions. In general, this also happens for the subsampling version (4.6) which has similar RB, RRMSE, and coverage as (4.1).

Finally, Table 1 suggests that, although the FPC corrections improves the Rao, Wu, and Yue (1992) estimator in terms of bias and stability, these artificial corrections are not always the best way to proceed; particularly, for situations where the second stage sampling may use small sampling fractions within certain clusters.

Table 1. RB, RRMSE, and Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{12}$, where $\mho(y_1) = 0.2965$ and $\mho(y_2) = 0.1951$.

m	n_I	n	f_I %	f %	RB% for eqs. (4.1)	(4.6)	(4.7)	(4.8)	RRMSE% for eqs. (4.1)	(4.6)	(4.7)	(4.8)	Coverage% for eqs. (4.1)	(4.6)	(4.7)	(4.8)
2	20	40	6	1	-4.4	-4.4	4.5	-1.5	46.1	46.4	51.5	48.4	91.9	91.8	92.9	92.1
	40	80	12	2	-1.7	-1.7	9.1	-3.6	31.9	32.9	38.0	32.9	93.4	93.3	94.5	93.1
	60	120	17	3	-1.8	-1.8	12.9	-6.8	25.3	27.2	33.5	26.4	93.8	93.7	95.3	93.1
	80	160	23	4	-1.3	-1.3	18.1	-9.3	21.0	24.4	32.9	23.1	94.2	94.0	96.0	93.1
	100	200	29	5	-0.6	-0.7	24.3	-11.8	18.3	23.4	35.3	21.6	94.4	94.0	96.5	92.7
	120	240	35	6	-0.8	-0.8	30.1	-15.1	16.0	23.1	38.5	21.8	94.4	94.1	97.0	92.4
	140	280	41	7	-0.6	-0.6	37.2	-18.5	14.4	23.9	43.7	23.0	94.7	94.2	97.5	91.9
	160	320	46	7	-0.5	-0.5	44.8	-22.4	13.1	25.5	50.1	25.4	94.7	94.0	97.9	91.3
4	20	80	6	2	-4.8	-4.8	5.5	-0.6	42.1	42.2	47.2	44.2	92.1	92.1	93.3	92.6
	40	160	12	4	-3.1	-3.1	10.6	-2.3	29.2	29.3	35.5	30.0	93.5	93.4	94.9	93.5
	60	240	17	6	-2.7	-2.7	16.6	-3.7	23.2	23.4	33.0	23.9	93.9	93.9	95.8	93.7
	80	320	23	7	-1.8	-1.8	24.7	-4.2	19.6	19.9	35.8	20.3	94.2	94.2	96.5	93.9
	100	400	29	9	-0.9	-0.9	34.1	-4.7	17.0	17.4	41.9	17.8	94.4	94.4	97.3	93.9
	120	480	35	11	0.0	0.0	45.1	-5.4	15.1	15.8	50.7	16.2	94.7	94.6	97.8	93.9
	140	560	41	13	0.0	0.0	56.4	-7.1	13.4	14.5	60.7	15.2	94.7	94.7	98.3	93.8
	160	640	46	15	-1.2	-1.2	67.5	-10.2	11.9	13.6	71.1	15.6	94.7	94.6	98.7	93.4
6	20	120	6	3	-4.9	-4.9	6.0	-0.1	40.1	40.1	45.0	42.0	92.1	92.1	93.4	92.7
	40	240	12	6	-3.0	-3.0	12.2	-0.8	27.7	27.7	34.4	28.5	93.5	93.5	95.1	93.7
	60	360	17	8	-1.3	-1.3	20.9	-0.1	22.1	22.1	34.6	22.7	94.1	94.1	96.2	94.2
	80	480	23	11	-1.3	-1.3	29.3	-0.7	18.4	18.5	38.3	18.9	94.3	94.3	96.9	94.4
	100	600	29	14	-0.8	-0.8	39.9	-0.7	16.2	16.3	46.2	16.6	94.5	94.5	97.6	94.5
	120	720	35	17	-1.4	-1.4	50.9	-1.6	14.2	14.3	55.5	14.6	94.5	94.4	98.1	94.4
	140	840	41	20	-0.2	-0.2	66.9	-0.8	12.6	12.9	70.4	13.0	94.6	94.6	98.6	94.5
	160	960	46	22	-1.1	-1.1	82.7	-2.0	11.3	11.7	85.4	11.7	94.5	94.5	99.0	94.4

5.2. Example 2: point estimator $\hat{R}_{34}$

The results for this example are summarised in Table 2. In terms of RB, it can be seen that the FPC-adjusted Rao, Wu, and Yue (1992) (4.8) estimator has the best performance when the sampling fractions are very small. This might be useful for highly stratified sampling designs. However, for increasing sampling fractions, the estimators (4.7) and (4.8) tend, respectively, to increasingly over- and under-estimate the variance. Again, this is more noticeable with small sample sizes at the second stage (small values of m).

On the other hand, in terms of RB, the variance estimators (4.1) and (4.6) tend consistently to zero for increasing sampling fractions, regardless of the utilised sample size at the second stage. This is something desirable in business surveys, for example, where sampling fractions are large or in situations

Table 2. RB, RRMSE, and Coverage at 95% confidence level of variance estimators for the point estimator $\hat{R}_{34}$, where $\mho(y_3) = 0.3181$ and $\mho(y_4) = 0.4958$.

m	n_I	n	f_I %	f %	RB% for eqs. (4.1)	(4.6)	(4.7)	(4.8)	RRMSE% for eqs. (4.1)	(4.6)	(4.7)	(4.8)	Coverage% for eqs. (4.1)	(4.6)	(4.7)	(4.8)
2	18	36	12	1	-3.5	-3.5	12.8	-0.8	42.2	43.0	51.7	44.1	93.1	93.0	94.8	93.3
	26	52	17	2	-2.4	-2.4	17.3	-3.0	33.6	35.1	44.9	34.4	93.7	93.6	95.7	93.6
	35	70	23	3	-1.7	-1.7	23.5	-5.3	27.8	30.5	43.2	28.3	94.1	93.9	96.4	93.5
	44	88	29	3	-1.4	-1.4	30.5	-7.8	23.8	27.8	44.9	24.6	94.3	94.0	97.0	93.4
	53	106	35	4	-1.1	-1.1	38.8	-10.3	20.9	27.0	49.6	22.5	94.4	94.1	97.5	93.2
	62	124	41	5	-1.1	-1.1	47.8	-13.3	18.5	26.8	56.2	21.8	94.5	94.0	98.0	92.8
	69	138	46	5	-0.9	-0.9	56.2	-15.7	17.0	27.7	63.1	22.1	94.6	94.0	98.3	92.4
4	18	72	12	3	-3.8	-3.8	14.1	0.4	39.9	39.9	49.5	41.7	92.9	92.9	94.8	93.4
	26	104	17	4	-2.6	-2.6	19.9	-0.9	31.9	32.0	44.2	32.6	93.6	93.6	95.8	93.8
	35	140	23	5	-2.0	-2.0	27.4	-2.3	26.4	26.6	44.2	26.6	94.0	94.0	96.6	94.0
	44	176	29	7	-1.7	-1.6	36.3	-3.7	22.7	22.9	48.2	22.7	94.2	94.2	97.3	93.9
	53	212	35	8	-1.4	-1.5	46.7	-5.1	19.8	20.3	55.4	20.0	94.4	94.3	97.9	93.9
	62	248	41	10	-1.2	-1.2	59.1	-6.7	17.6	18.3	65.7	18.2	94.5	94.5	98.4	93.8
	69	276	46	11	-1.0	-1.0	70.4	-8.0	16.2	17.1	75.9	17.2	94.5	94.5	98.7	93.6
6	18	108	12	4	-4.1	-4.1	14.6	0.8	38.9	38.9	48.5	40.8	92.9	92.9	94.9	93.5
	26	156	17	6	-2.9	-2.9	20.7	-0.2	31.1	31.1	43.8	31.9	93.6	93.6	95.9	93.9
	35	210	23	8	-1.9	-1.9	29.3	-0.8	25.8	25.8	44.9	26.1	94.0	94.0	96.8	94.1
	44	264	29	10	-1.6	-1.6	39.0	-1.7	22.1	22.2	50.0	22.1	94.2	94.2	97.4	94.2
	53	318	35	12	-1.7	-1.7	49.9	-3.0	19.3	19.4	57.9	19.2	94.3	94.3	98.0	94.2
	62	372	41	15	-1.5	-1.5	63.7	-4.0	17.1	17.2	69.7	17.0	94.4	94.4	98.5	94.1
	69	414	46	16	-1.0	-1.0	77.0	-4.4	15.7	15.9	81.8	15.7	94.5	94.5	98.9	94.1

where stratification is moderate. Note that the RB for the proposed estimators always showed a slight negative bias. This is something expected and well-documented when using the Hájek (1964) approximations (see Haziza, Mecatti, and Rao (2008); Brewer and Donadio (2003)).

As to stability of the studied variance estimators, Table 2 shows that the proposed estimator (4.1) has the smallest RRMSE in all considered situations. It can also be seen that the subsampling version (4.6) has small but slightly higher RRMSE. In terms of coverage, it can be seen that the estimator (4.7) has better coverage than the FPC-adjusted (4.8). The coverage of the proposed estimators (4.1) and (4.6) become closer to 95% for increasing sampling fractions. Overall, the worst coverage was showed by the estimator (4.8). This suggests again that the fixed *ad hoc* FPC correction might not be suitable.

6. Conclusion

Self-weighted two-stage sampling designs are common in practice. Besides

their popularity, it is also common practice to compute variance estimates relying only on the first sampling stage (e.g., Särndal, Swensson, and Wretman (1992, Chap. 4)). A customary jackknife variance estimator for sampling designs of more than one stage is the Rao, Wu, and Yue (1992) estimator, originally designed for functions of totals and for negligible sampling fractions. This jackknife would work well when most of the variability is between clusters and with very small sampling fractions (highly stratified samples), but this may not necessarily be the case.

First, we propose an alternative Sen-Yates-Grundy form of the generalised unequal-probability without-replacement jackknife variance estimator (Campbell (1980)). This estimator is extended to two-stage sampling by proposing new less restrictive regularity conditions than those from Berger and Skinner (2005), thus allowing two-stage sampling for the Horvitz-Thompson (original) form of the Campbell (1980) generalised jackknife as well.

Secondly, we propose a novel design-consistent jackknife variance estimator for self-weighted two-stage without-replacement sampling. The proposed estimator does not need joint-inclusion probabilities, allows stratification, naturally includes FPC, and comprises a wide class of point estimators (functions of means). Monte-Carlo simulations show that the proposed estimator can be more accurate than customary jackknife estimators, specially in situations where the first stage sampling fraction is large, or in cases where the second stage sampling fractions are small. The proposed estimator incorporates not only clustering effects but also the underlying unequal-probabilities of both, clusters and observations.

Acknowledgements

The authors are grateful to Chris J. Skinner (London School of Economics, U.K.), Nikolaos Tzavidis (University of Southampton, U.K.), Ray L. Chambers (University of Wollongong, Australia), and three anonymous referees for helpful comments. Emilio López Escobar was supported by the CONACyT (Mexico).

Appendix

A. Proof of Theorem 1

The proof uses standard arguments in proving jackknife variance estimators design consistency (see Miller (1964); Shao and Tu (1995, Sec. 2.1.1)). Hence, from the mean value theorem we have that $\hat{\theta}-\hat{\theta}_{(k)} = g(\hat{\boldsymbol{\mu}})-g(\hat{\boldsymbol{\mu}}_{(k)}) = \boldsymbol{\nabla}(\boldsymbol{\xi}_k)^T(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}}_{(k)}) = \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}}_{(k)}) + r_k^*$, where $\boldsymbol{\xi}_k$ denotes a point between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}_{(k)}$, and where $r_k^* = (\boldsymbol{\nabla}(\boldsymbol{\xi}_k) - \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}))^T(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}}_{(k)})$ is the remainder term. Thus, $\varepsilon_{(k)} = \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T(1-\tilde{w}_k)(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}}_{(k)}) + r_k$, where

$$r_k = (1-\tilde{w}_k)r_k^*. \tag{A.1}$$

It can be shown that

$$(1-\tilde{w}_k)(\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}_{(k)})=\tilde{w}_k(\boldsymbol{y}_k-\tilde{\boldsymbol{\mu}}), \tag{A.2}$$

implying that

$$\varepsilon_{(k)}=\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T\tilde{w}_k(\boldsymbol{y}_k-\hat{\boldsymbol{\mu}})+r_k. \tag{A.3}$$

Furthermore, the Cauchy inequality together with (A.1) and (A.2) imply

$$|r_k|\leq||\boldsymbol{\nabla}(\boldsymbol{\xi}_k)-\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||\tilde{w}_k||\boldsymbol{y}_k-\hat{\boldsymbol{\mu}}||. \tag{A.4}$$

The regularity condition C7 implies that there are constants $\lambda>0$, δ, and $0\leq\beta<1$, where $\beta/2<\delta$, such that

$$||\boldsymbol{\nabla}(\boldsymbol{\xi}_k)-\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||\leq\lambda||\boldsymbol{\xi}_k-\hat{\boldsymbol{\mu}}||^{\delta}. \tag{A.5}$$

As $\boldsymbol{\xi}_k$ is between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}_{(k)}$, we have that $||\boldsymbol{\xi}_k-\hat{\boldsymbol{\mu}}||\leq||\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\mu}}_{(k)}||$. Combining this with (A.2), we obtain $||\boldsymbol{\xi}_k-\hat{\boldsymbol{\mu}}||\leq||(1-\tilde{w}_k)^{-1}\tilde{w}_k(\boldsymbol{y}_k-\hat{\boldsymbol{\mu}})||$, which by (A.5) gives $||\boldsymbol{\nabla}(\boldsymbol{\xi}_k)-\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||\leq\lambda|1-\tilde{w}_k|^{-\delta}\tilde{w}_k^{\delta}||\boldsymbol{y}_k-\hat{\boldsymbol{\mu}}||^{\delta}$. Then, by C2 this becomes $||\boldsymbol{\nabla}(\boldsymbol{\xi}_k)-\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||\leq\lambda\alpha^{-\delta}\tilde{w}_k^{\delta}||\boldsymbol{y}_k-\hat{\boldsymbol{\mu}}||^{\delta}$ which, combined with (A.4), imply

$$|r_k|\leq\lambda\alpha^{-\delta}\tilde{w}_k^{1+\delta}||\boldsymbol{y}_k-\hat{\boldsymbol{\mu}}||^{1+\delta}. \tag{A.6}$$

Moreover, C3 implies that

$$\{n\ \mathrm{var}(\hat{\theta})_L\}^{-2}=\mathcal{O}(1). \tag{A.7}$$

By substituting (A.3) in (3.3), we obtain $\widehat{\mathrm{var}}(\hat{\theta})_{SYG}=A+2(E-C)+D-B$, where

$$\begin{aligned}
A&=\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})^T\ \widehat{\mathbf{var}}(\hat{\boldsymbol{\mu}})_{SYG}\ \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}),\\
B&=\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ r_\ell,\\
C&=\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ \tilde{w}_\ell(\boldsymbol{y}_\ell-\hat{\boldsymbol{\mu}})^T\ \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}),\\
D&=\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k^2,
\end{aligned} \tag{A.8}$$

$$E=\sum_{k\in s}\sum_{\ell\in s}\mathcal{D}_{k\ell}\ r_k\ \tilde{w}_k(\boldsymbol{y}_k-\hat{\boldsymbol{\mu}})^T\ \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}). \tag{A.9}$$

Hence, Theorem 1 follows if we show

$$\frac{A}{\text{var}(\hat{\theta})_L} \to_p 1, \tag{A.10}$$

$$\frac{B}{\text{var}(\hat{\theta})_L} \to_p 0, \tag{A.11}$$

$$\frac{C}{\text{var}(\hat{\theta})_L} \to_p 0, \tag{A.12}$$

$$\frac{D}{\text{var}(\hat{\theta})_L} \to_p 0, \tag{A.13}$$

$$\frac{E}{\text{var}(\hat{\theta})_L} \to_p 0, \tag{A.14}$$

Now C1 implies (A.10), whereas (A.11) and (A.12) can be shown by following the Berger and Skinner (2005) proof and taking into account the changes in regularity conditions C5 to C7. Hence, it remains to show (A.13) and (A.14). We start with (A.13). By the triangle and the Cauchy inequalities, (A.8) implies

$$|D| \leq \sum_{k\in s}\sum_{\ell\in s} |\mathcal{D}_{k\ell}||r_k|^2 = D_1 + D_2 \leq (G_s^{1/2} + H_s^{1/2})D_3^{1/2},$$

where $D_1 = \sum\sum_{k,\ell\in s} \mathcal{D}^-_{k\ell}|r_k|^2 \leq G_s^{1/2}D_3^{1/2}$, $D_2 = \sum\sum_{k,\ell\in s} \mathcal{D}^+_{k\ell}|r_k|^2 \leq H_s^{1/2}D_3^{1/2}$ and

$$D_3 = n^\beta \sum_{k\in s}\sum_{\ell\in s} |r_k|^4 = n^{1+\beta}\sum_{k\in s}|r_k|^4. \tag{A.15}$$

Thus, (A.13) follows from C5 and C6 if we show $D_3\{\text{var}(\hat{\theta})_L\}^{-2} \to_p 0$. Using the expression (A.6) in (A.15), we have that

$$\frac{D_3}{\text{var}(\hat{\theta})_L^2} \leq \frac{\lambda^4}{\alpha^{4\delta}} \frac{n^{4+\beta}}{\{n\ \text{var}(\hat{\theta})_L\}^2} \left(\frac{1}{n}\sum_{k\in s} \tilde{w}_k^{4(1+\delta)} ||\boldsymbol{y}_k - \hat{\boldsymbol{\mu}}||^{4(1+\delta)}\right). \tag{A.16}$$

Conditions C3, C4, and (A.7) and (A.16) imply $D_3\{\text{var}(\hat{\theta})_L\}^{-2} = n^\beta \mathcal{O}_p(n^{-4\delta})$. From C7, $\beta < 4\delta$. Thus $D_3\{\text{var}(\hat{\theta})_L\}^{-2} \to_p 0$, implying (A.13). We now show (A.14). Using the triangle and the Cauchy inequalities in (A.9) gives

$$|E| \leq \sum_{k\in s}\sum_{\ell\in s} |\mathcal{D}_{k\ell}||r_k||\tilde{y}_k| = E_1 + E_2 \leq (G_s^{1/2} + H_s^{1/2})E_3^{1/2},$$

where $E_1 = \sum\sum_{k,\ell\in s} \mathcal{D}^-_{k\ell}|r_k||\tilde{y}_k|$, $E_2 = \sum\sum_{k,\ell\in s} \mathcal{D}^+_{k\ell}|r_k||\tilde{y}_k|$, and

$$E_3 = n^\beta \sum_{k\in s}\sum_{\ell\in s} |r_k|^2|\tilde{y}_k|^2 = n^{1+\beta}\sum_{k\in s}|r_k|^2|\tilde{y}_k|^2, \tag{A.17}$$

$$\tilde{y}_k = \tilde{w}_k(\boldsymbol{y}_k - \hat{\boldsymbol{\mu}})^T \boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}). \tag{A.18}$$

Thus, (A.14) follows from C5 and C6 if we show $E_3\{\mathrm{var}(\hat{\theta})_L\}^{-2} \to_p 0$. Using the Cauchy inequality in (A.18), we have that $|\tilde{y}_k| \le \tilde{w}_k||\boldsymbol{y}_k - \hat{\boldsymbol{\mu}}||\ ||\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||$. This inequality together with (A.6) and (A.17) imply that

$$\frac{E_3}{\mathrm{var}(\hat{\theta})_L^2} \le ||\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}})||^2 \frac{\lambda^2}{\alpha^{2\delta}} \frac{n^{4+\beta}}{\{n\ \mathrm{var}(\hat{\theta})_L\}^2} \left(\frac{1}{n} \sum_{k \in s} \tilde{w}_k^{4+2\delta} ||\boldsymbol{y}_k - \hat{\boldsymbol{\mu}}||^{4+2\delta} \right). \quad \text{(A.19)}$$

From C7, $\beta < 2\delta$. This, together with (A.19), (A.7), C4, and C8 imply $E_3\{\mathrm{var}(\hat{\theta})_L\}^{-2} \to_p 0$, completing the proof.

B. Proof of Theorem 2

Theorem 1 asserts the consistency of the variance estimator $\widehat{\mathrm{var}}(\hat{\theta})_{SYG}$ at (3.3), used to develop the proposed variance estimator $\widehat{\mathrm{var}}(\hat{\theta})_{prop}$ from (4.1). Hence, given conditions C1, C3, C4, C7, and C8, it remains to show that $M_i \ge m \ge 2$, for all $i = 1, \ldots, N_I$, implies that C2, C5, and C6 hold. From self-weighting, it can be shown that C2 holds. We now consider C5 and C6. Let $q_{Ii} = 1 - \pi_{Ii}$ and $q^*_{Ii} = 1 - \pi^*_{Ii}$, and let $\beta = \mathrm{Log}(n_I)/\mathrm{Log}(n) < 1$ be such that $n^\beta = n_I$. It can be shown that $|q^*_{Ii}| = \mathcal{O}(1)$ for $M_i \ge m \ge 2$, for all $i = 1, \ldots, N_I$, and that $d = \mathcal{O}(N_I)$ as $q_{Ii} = \mathcal{O}(1)$.

If $q^*_{Ii} > 0$, we have from C5 and C6 that $\mathcal{D}^-_{k\ell} = q_{Ii}\ q_{Ij}/d = \mathcal{O}_p({N_I}^{-1})$ for $k \in s_i, \ell \in s_j, i \ne j$, and that $\mathcal{D}^+_{k\ell} = q^*_{Ii}$ for $(k \ne \ell) \in s_i$. Thus,

$$G_s = \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1, i \ne j}^{n_I} \sum_{k=1}^{m} \sum_{\ell=1}^{m} \left(\frac{q_{Ii}\ q_{Ij}}{d} \right)^2 = \frac{m^2 {n_I}^2}{n_I {N_I}^2} \mathcal{O}_p(1) = f_I^2 m^2 \mathcal{O}_p(n_I^{-1}),$$

$$H_s = \frac{1}{n_I} \sum_{i=1}^{n_I} \sum_{j=1, i=j}^{n_I} \sum_{k=1}^{m} \sum_{\ell=1, k \ne \ell}^{m} (q^*_{Ii})^2 = m(m-1)\mathcal{O}_p(1),$$

where $f_I = n_I/N_I$ and m are constants. Moreover, if $\hat{d}$ was used instead of d in (4.2), then $G_s = m^2 \mathcal{O}_p(n_I^{-1})$. Now, if $q^*_{Ii} < 0$, we have that $\mathcal{D}^-_{k\ell} = q_{Ii}\ q_{Ij}/d = \mathcal{O}_p({N_I}^{-1})$ for $k \in s_i, \ell \in s_j, i \ne j$, $\mathcal{D}^-_{k\ell} = q^*_{Ii}$ for $(k \ne \ell) \in s_i$, and $\mathcal{D}^+_{k\ell} = 0$. Hence, $G_s = f_I^2 m^2 \mathcal{O}_p(n_I^{-1}) + m(m-1)\mathcal{O}_p(1)$ and $H_s = 0$. Again, if $\hat{d}$ was used instead of d, $G_s = m^2 \mathcal{O}_p(n_I^{-1}) + m(m-1)\mathcal{O}_p(1)$.

Thus, G_s and H_s are $\mathcal{O}_p(1)$ completing the proof.

C. Proof of (4.1)

We need some results and approximations. For the designs and estimators defined in Sections 2 and 4, we have that

$$\varsigma_{(Ii)} = \sum_{k \in s_i} \varepsilon_{(k)}. \quad \text{(A.1)}$$

We show (A.1) by recalling that the sampling design is self weighted. Hence, $n = n_I m$, $\tilde{w}_k = n^{-1}$, $\tilde{w}_{\ell(k)} = (n-1)^{-1}$, and $\tilde{w}_{k(Ii)} = (n-m)^{-1}$, implying

$$\begin{aligned}\frac{n_I - 1}{n_I}(\hat{\mu}_q - \hat{\mu}_{q(Ii)}) &= \frac{1}{n_I}\Big[(n_I-1)\hat{\mu}_q - \frac{n_I - 1}{m(n_I-1)}\Big(n_I m\hat{\mu}_q - \sum_{k\in s_i} y_{qk}\Big)\Big] \\ &= \sum_{k\in s_i}\frac{(n-1)\hat{\mu}_q - n\hat{\mu}_q + y_{qk}}{n} = \frac{n-1}{n}\sum_{k\in s_i}(\hat{\mu}_q - \hat{\mu}_{q(k)}).\end{aligned}$$

Then, as $g(\cdot)$ is linearisable, we obtain (A.1). Now, under the asymptotic framework from Subsection 3.1 (Isaki and Fuller (1982)) we have that $N \to \infty$, implying $N_I \to \infty$ as m is assumed fixed. Then we have that $d \to \infty$ and $q_{Ii} = \mathcal{O}(1)$. Letting $q_{Ii} = 1 - \pi_{Ii}$ and $q^*_{Ii} = 1 - \pi^*_{Ii}$, we introduce the approximations that are suitable for large values of d:

$$q_{Ii}\ q_{Ij}\ \{q_{Ii}\ q_{Ij} - d\}^{-1} \doteq -\frac{q_{Ii}q_{Ii}}{d}, \tag{A.2}$$

$$\frac{(\hat{d} - q_{Ii})}{d} \doteq 1, \tag{A.3}$$

$$q^*_{Ii} + \frac{q^2_{Ii}}{d} \doteq q^*_{Ii}. \tag{A.4}$$

Hence, from (3.3), we have that

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\theta})_{SYG} &= \sum_{k\in s}\sum_{\ell\in s,\ell\neq k}\mathcal{D}_{kl}\ \varepsilon_{(k)}\varepsilon_{(\ell)} - \sum_{k\in s}\sum_{\ell\in s,\ell\neq k}\mathcal{D}_{kl}\ \varepsilon^2_{(k)} \\ &= \sum_{i\in s}\sum_{j\in s,j=i}\sum_{k\in s_i}\sum_{\ell\in s_j,\ell\neq k}\mathcal{D}_{kl}\ \varepsilon_{(k)}\varepsilon_{(\ell)} + \sum_{i\in s}\sum_{j\in s,j\neq i}\sum_{k\in s_i}\sum_{\ell\in s_j}\mathcal{D}_{kl}\ \varepsilon_{(k)}\varepsilon_{(\ell)} \\ &\quad - \sum_{i\in s}\sum_{j\in s,j=i}\sum_{k\in s_i}\sum_{\ell\in s_j,\ell\neq k}\mathcal{D}_{kl}\ \varepsilon^2_{(k)} - \sum_{i\in s}\sum_{j\in s,j\neq i}\sum_{k\in s_i}\sum_{\ell\in s_j}\mathcal{D}_{kl}\ \varepsilon^2_{(k)}.\end{aligned}$$

Then, substituting for $\mathcal{D}_{kl}$ (see Section 4), and using (A.1) and (A.2), we obtain

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\theta})_{SYG} &\doteq \sum_{i\in s} q^*_{Ii}\Big[\varsigma^2_{(Ii)} - \sum_{k\in s_i}\sum_{\ell\in s_i\ell=k}\varepsilon_{(k)}\varepsilon_{(\ell)}\Big] - \sum_{i\in s}\sum_{j\in s,j\neq i}\frac{q_{Ii}\ q_{Ij}}{d}\varsigma_{(Ii)}\varsigma_{(Ij)} \\ &\quad -(m-1)\sum_{i\in s} q^*_{Ii}\sum_{k\in s_i}\varepsilon^2_{(k)} + m\sum_{i\in s}\sum_{j\in s,j\neq i}\frac{q_{Ii}\ q_{Ij}}{d}\sum_{k\in s_i}\varepsilon^2_{(k)} \\ &= \sum_{i\in s}\Big[q^*_{Ii} + \frac{q^2_{Ii}}{d}\Big]\varsigma^2_{(Ii)} - \frac{1}{d}\Big[\sum_{i\in s} q_{Ii}\varsigma_{(Ii)}\Big]^2 + m\sum_{i\in s}\Big[q_{Ii}\frac{\widehat{d} - q_{Ii}}{d} - q^*_{Ii}\Big]\sum_{k\in s_i}\varepsilon^2_{(k)}.\end{aligned}$$

Now, by using (A.3) and (A.4), combined with $q_{Ii} - q^*_{Ii} = (\pi_{Ii} - f)/(m-1) = \pi_{Ii}(M_i - m)/M_i(m-1)$, we obtain the estimator $\widehat{\mathrm{var}}(\hat{\theta})_{prop}$ from (4.1).

References

Berger, Y. G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953-964.

Berger, Y. G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pak. J. Statist.* **27**, 407-426.

Berger, Y. G. and Rao, J. N. K. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *J. Roy. Statist. Soc. Ser. B* **68**, 531-547.

Berger, Y. G. and Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *J. Roy. Statist. Soc. Ser. B* **67**, 79-89.

Berger, Y. G. and Tillé, Y. (2009). Sampling with unequal probabilities. In *Sample Surveys: Design, Methods and Applications* (Edited by D. Pfeffermann and C. R. Rao), 39-54. Elsevier, Amsterdam.

Brewer, K. R. W. (1975). A simple procedure for πpswor. *Austral. J. Statist.* **17**, 166-172.

Brewer, K. R. W. and Donadio, M. E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodol.* **29**, 189-196.

Brewer, K. R. W. and Gregoire, T. G. (2009). Introduction to survey sampling. In *Sample Surveys: Design, Methods and Applications* (Edited by D. Pfeffermann and C. R. Rao), 9-37. Elsevier, Amsterdam.

Campbell, C. (1980). A different view of finite population estimation. *Proc. Surv. Res. Meth. Sect., Amer. Statist. Assoc.* 319-324.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* **35**, 1491-1523.

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference* (Edited by V. P. Godambe and D. A. Sprott), p.236. Holt, Rinehart and Winston, Toronto.

Hájek, J. (1981). *Sampling From a Finite Population.* Dekker, New York.

Haziza, D., Mecatti, F. and Rao, J. N. K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron* **LXVI**, 91-108.

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

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77**, 89-96.

Kim, J. K. and Sitter, R. R. (2003). Efficient replication variance estimation for two-phase sampling. *Statist. Sinica* **13**, 641-653.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *J. Roy. Statist. Soc. Ser. B* **36**, 1-37.

Kovar, J. G., Rao, J. N. K. and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canad. J. Statist.* **16**, 25-45.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9**, 1010-1019.

Miller, R. G. (1964). A trustworthy jackknife. *Ann. Math. Statist.* **35**, 1594-1605.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353-360.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Surv. Methodol.* **18**, 209-217.

Robinson, P. M. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: Indian J. Statist. Ser. B* **45**, 240-248.

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

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agr. Statist.* **5**, 119-127.

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

Tillé, Y. (2006). *Sampling Algorithms.* Springer, New York.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist.* **29**, 614.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley, New York.

Wolter, K. M. (2007). *Introduction to Variance Estimation.* 2nd edition. Springer, New York.

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.

Departamento de Estadística, ITAM, Río Hondo 1, Tizapán San Angel, 01000 Distrito Federal, México.

E-mail: Emilio.Lopez@itam.mx

Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton. SO17 1BJ, United Kingdom.

E-mail: Y.G.Berger@soton.ac.uk

(Received October 2011; accepted May 2012)