Statistica Sinica 8(1998), 1131-1151

VARIANCE ESTIMATION FOR SUPERPOPULATION PARAMETERS

Edward L. Korn and Barry I. Graubard

National Cancer Institute

Abstract: In scientific applications, interest usually focuses on the "superpopulation" parameters of a stochastic model hypothesized to underlie the generation of the values in a finite population, rather than finite-population parameters themselves. Variance formulas for sampled data that incorporate finite-population correction factors are not appropriate for these applications. For simple random sampling, it is common practice to ignore these correction factors in variance estimation; this yields correct superpopulation inference under a simple superpopulation model. This is shown to hold true for two-stage simple random sampling of clusters, but not for stratified sampling or probability-proportional-to-size sampling. Asymptotically unbiased variance estimators are provided for these latter two types of sampling that are appropriate for superpopulation inference under a general superpopulation model. An application is given using data from the 1987 National Health Interview Survey which shows that the difference between classical repeated-sampling variance estimators and a superpopulation variance estimator can be quite large.

Key words and phrases: Cluster sampling, finite-population correction factors, probability-proportional-to-size sampling, stratified sampling.

1. Introduction

Classical sampling theory concerns inference for finite population parameters, e.g., the mean of all the values of a variable Y over the units in the target population. Stochastic models for Y, also known as superpopulation models (Deming and Stephan (1941)) have been used extensively to evaluate designs and estimators (Cochran (1946), Hartley and Sielken (1975), Cassel, Sarndal and Wretman (1977), ch. 4-6), to estimate means for small areas (Ghosh and Rao (1994)) to incorporate measurement error (Sarndal, Swensson and Wretman (1992), ch. 16), and to handle missing data (Little and Rubin (1987), ch. 12). The parameters of the stochastic models themselves, however, are probably of more interest than the finite-population parameters for studies involving questions of science (as opposed to administrative or quality assurance applications). Cochran (1977), p. 39 and Yates (1981), p. 178 suggest that in comparing two domain means with simple random sampling that the finite-population correction factors should be ignored, since interest will usually be in the superpopulation

means (see also Deming (1953)). This advice is easily justified in section 2 below using a simple superpopulation model.

In this paper we investigate variance estimation for superpopulation parameters under some more general without-replacement sampling designs. We ask whether it is appropriate to use with-replacement variance estimators that ignore finite-population correction factors for these designs. To address this, we utilize a very general superpopulation model in section 3 that includes clustering, and for pedagogical reasons, a simpler superpopulation model in section 2. We find that for two-stage sampling with simple random sampling (without replacement) at the first stage, the with-replacement variance estimator is appropriate. For stratified simple random sampling at the first stage, however, an adjustment to the with-replacement variance estimator is required. For probability-proportional-to-size (pps) sampling from within strata, the with-replacement estimator is not easily modified to achieve consistent estimation of the superpopulation variance. In this situation, a modification of the without-replacement variance estimator is given, and an additional ad hoc modification of the with-replacement estimator is given for cases in which the joint inclusion probabilities are not known to the analyst. The robustness of the variance estimation to misspecification of the superpopulation model is considered.

In section 4 we present an application using data from the 1987 National Health Interview Survey which shows that use of classical design-based variance estimation can drastically underestimate the variance appropriate for a superpopulation parameter. We restrict attention to variance estimation of the superpopulation mean in sections 2 and 3, and consider other parameters in the Discussion. We also give, in the Discussion, our reasons for why we believe that the superpopulation variance estimators presented in this paper are often more appropriate than classical repeated-sampling variance estimators. Sufficient conditions for the asymptotic results are given in Appendix B; derivations use standard conditioning and Taylor series arguments and are omitted for the most part.

2. Unclustered Finite-Population/Superpopulation Model

The finite population consists of $(Y_1, \eta_1), \ldots, (Y_K, \eta_K)$ with Y_i being a realization of a random variable with mean m_i and variance t_i^2, and with $Y_1, \ldots, Y_K$ being independent. The t_i^2 represents the measurement error variance, and η_i represents a stratum variable that can be used for stratified sampling. The (m_i, t_i^2, η_i) are assumed to be independent and identically distributed (i.i.d.), each with the same distribution as the random vector (μ, σ^2, η) which has trivariate distribution F. The finite-population mean is always defined as the simple mean of the Y observations in the finite population. The superpopulation mean is

defined as $\mu_{SP} = E_F(\mu)$. Note that unlike Potthoff et al. (1992), the superpopulation mean does not depend upon the realized finite population or the realized sample from the finite population (see also Kott (1993)). With no measurement error and no stratum variable, the model considered here is the superpopulation model considered by Koop (1985).

Case 2.1. Simple random sampling without replacement

Let $y_1, \ldots, y_k$ be the values sampled from the finite population, and $\bar{y}$ be the sample mean. An unbiased estimator of the repeated-sampling variance of $\bar{y}$ is given by

$$\hat{\text{Var}}_{wor}(\bar{y}) = \frac{(1-f)}{k} \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2, \tag{2.1}$$

where $(1-f) = (K-k)/K$ is the finite-population correction factor. If the finite-population correction factor is set to 1, then one obtains the repeated-sampling formula that would have been used if the sampling had actually been simple random sampling with replacement:

$$\hat{\text{Var}}_{wr}(\bar{y}) = \frac{1}{k} \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2. \tag{2.2}$$

Using the superpopulation model, we have $\text{Var}\,(\bar{y}) = E\{\hat{\text{Var}}_{wr}(\bar{y})\} = [E_F(\sigma^2) + \text{Var}_F(\mu)]/k$, and $E[\hat{\text{Var}}_{wor}(\bar{y})] = (1-f)[E_F(\sigma^2) + \text{Var}_F(\mu)]/k$, confirming the advice to ignore the finite-population correction factor for superpopulation inference (see also Fuller (1975)). (Expectations and variances of sampled quantities are to be interpreted as including the randomness from both the generation of the finite population (using the superpopulation model) and the sampling of the finite population.)

Case 2.2. Stratified simple random sampling without replacement

Assume that there are L strata, and let $\eta \in \{1, \ldots, L\}$ indicate in which of the strata an observation appears. Let K be the number of observations in the finite population and K_h be the number of observations in the finite population in stratum h. From stratum h, $k_h = c_h(K_h)$ observations are sampled as a simple random sample without replacement: $y_{h1}, \ldots, y_{hk_h}$. The functions c_h can depend upon h since we may wish to utilize different sampling rates depending upon prior knowledge of stratum characteristics, e.g., the variability of the y's in the different strata. Let $\bar{y} = (K_1\bar{y}_1 + \cdots + K_L\bar{y}_L)/K$ be the stratified mean, where $\bar{y}_h$ is the mean of the sampled observations in stratum h. The repeated-sampling

variance estimator, ignoring the finite-population correction factors, is given by

$$\hat{\mathrm{Var}}_{wr}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{1}{k_h} s_h^2,$$

where

$$s_h^2 = \frac{1}{k_h - 1} \sum_{i=1}^{k_h} (y_{hi} - \bar{y}_h)^2.$$

(By the "repeated-sampling variance" of a sampled quantity, we mean the variance of that quantity over repeated independent samples from a fixed finite population.) Using the superpopulation model, we have

$$\mathrm{Var}\,(\bar{y}) = \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E\Big[\frac{K_h^2}{c_h(K_h)}\Big] + \Delta_{st} \qquad \text{and}$$

$$E[\hat{\mathrm{Var}}_{wr}(\bar{y})] = \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E\Big[\frac{K_h^2}{c_h(K_h)}\Big],$$

where $\Delta_{st} = \frac{1}{K}[\sum_{h=1}^{L} \pi_h \mu_h^2 - (\sum_{h=1}^{L} \pi_h \mu_h)^2]$, $\pi_h = P_F[\eta = h]$, $\mu_h = E_F[\mu|\eta = h]$, and $\sigma_h^2 = \mathrm{Var}_F[\mu|\eta = h] + E_F[\sigma^2|\eta = h]$. To get a feel for the magnitude of the underestimation of the variance, Table 1 presents the relative bias of $\hat{\mathrm{Var}}_{wr}(\bar{y})$ when the sampling fractions are the same for the different strata. In this table, the "between-strata variance" refers to $K\Delta_{st}$, and the "within-stratum variance" refers to $\sum \pi_h \sigma_h^2$. Recall that we are considering the performance of the variance estimator without finite-population correction factors. For stratified variance estimators with the correction factors, the relative biases in Table 1 would be larger. With or without finite-population correction factors, however, the relative bias of $\hat{\mathrm{Var}}_{wr}(\bar{y})$ is negligible with small sampling fractions.

Table 1. Relative (negative) bias of the variance estimator $\hat{\mathrm{Var}}_{wr}(\bar{y})$ for stratified sampling with the same sampling fractions in the different strata

	Ratio of between-strata to within-stratum variances		
Sampling Fraction	0.1	1	2
1%	< 1%	1%	2%
10%	1%	9%	17%
25%	2%	20%	33%

For large sampling fractions, one should correct $\hat{\mathrm{Var}}_{wr}(\bar{y})$ for its underestimation. This can be done by adding the following term to $\hat{\mathrm{Var}}_{wr}(\bar{y})$, which is an

unbiased estimator of Δ_{st}:

$$\hat{\Delta}_{st} = -\sum_{h=1}^{L} \frac{K_h(K-K_h)}{K^2(K-1)} \frac{1}{k_h} s_h^2 + \frac{1}{K-1}\Big[\sum_{h=1}^{L} \frac{K_h}{K}\bar{y}_h^2 - \bar{y}^2\Big].$$

The unbiased estimator of the variance of $\bar{y}$ as an estimator of μ_{SP} is then given by

$$\hat{\mathrm{Var}}_{SP}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h(K_h-1)}{K(K-1)} \frac{1}{k_h} s_h^2 + \frac{1}{K-1}\Big[\sum_{h=1}^{L} \frac{K_h}{K}\bar{y}_h^2 - \bar{y}^2\Big].$$

The question arises as to whether there is a constraint that the strata used for the sampling must be the same as the strata defined in the superpopulation. The answer is no. The strata being used for the sampling are always determined by some unit level characteristics. Whatever these characteristics are, we can typically imagine them being incorporated into the variable η defined in the superpopulation. However, it is possible that some information about the realized finite population is also used in defining the strata. For example, the units in the finite population may be divided into strata of exactly size 1000 by counting off groups of 1000 units based on an ordering of a continuous variable observed on the finite-population units. The η notation used above does not strictly cover this case, but can be generalized to handle it by allowing η to be continuous and the stratum of unit i to be determined as a function of η_i and the finite-population values $(\eta_1, \ldots, \eta_K)$. The unbiasedness of $\hat{\mathrm{Var}}_{SP}(\bar{y})$ holds generally; the proof is given in Appendix A. For the later cases considered in this paper, we will use the simpler notation in which η_i is the stratum variable, although the results apply to the more general situation.

An alternative approach to using $\hat{\mathrm{Var}}_{SP}(\bar{y})$ would be to use an unstratified with-replacement variance estimator. For example, a referee suggests consideration of the following variance estimator, which treats the sample as if it had been a with-replacement probability-proportional-to-size sample: $\hat{v}(\bar{y}) = \Sigma\Sigma(z_{hi} - \bar{z})^2/k(k-1)$, where $k = \Sigma k_h$, $z_{hi} = K_h k y_{hi}/(K k_h)$, and $\bar{z} = \Sigma\Sigma z_{hi}/k (= \bar{y})$. However, note that $\hat{v}(\bar{y}) > 0$ even if $Y \equiv 1$ (so that $y_{hi} \equiv 1$ and $\mathrm{Var}(\bar{y}) = 0$), showing that $\hat{v}(\bar{y})$ is a biased estimator.

3. Clustered Finite-Population/Superpopulation Model

The finite population consists of K primary clusters. The ith primary cluster contains N_i secondary clusters. For the jth secondary cluster of the ith primary cluster, we assume that there are M_{ij} population values (Y's) and totals $T_{ij} = Y_{ij1} + \cdots + Y_{ijM_{ij}}$. The ith primary cluster also has a stratum variable η_i and a "size" variable Z_i which will be useful for probability-proportional-to-size (pps) sampling. We assume that the (M_{ij}, T_{ij}) are i.i.d. from a distribution with

mean $(\mathcal{M}_i, \mathcal{J}_i)$ and covariance matrix φ_i , and that $(\mathcal{M}_i, \mathcal{J}_i, \varphi_i, N_i, Z_i, \eta_i)$ are i.i.d. random variables with distribution function G. The superpopulation mean is defined by $\mu_{SP} = E_G(N\mathcal{J})/E_G(N\mathcal{M})$.

To avoid notational meltdown, we have not explicitly modeled measurement error for the Y's as we did in section 2; all the results of this section go through with measurement error added. Additionally, for pedagogical reasons, it will be useful in cases 3.1 – 3.4 to use a special case of the above model in which the secondary clusters are the final units. In this case we will use the following notation: within a primary cluster, Y_{ij} are i.i.d. with mean m_i and variance $t_i^2, j = 1, \ldots, N_i$; across the primary clusters, $(m_i, t_i^2, N_i, Z_i, \eta_i)$ are i.i.d random variables, each with the same distribution as the random vector $(\mathcal{M}, \sigma^2, N, Z, \eta)$ which has distribution function G. For this reduced model, the superpopulation mean is $\mu_{SP} = E_G(N\mu)/E_G(N)$.

Case 3.1. Two-stage sampling using simple random sampling without replacement

At the first stage of sampling, a simple random sample without replacement of k primary clusters is selected. At the second stage of sampling, a simple random sample without replacement of size $n_i = g(N_i)$ is selected. For example, $n_i \equiv n$ represents equal cluster sample sizes, and $n_i = \zeta N_i$ represents a self-weighting design.

As an estimator of μ_{SP}, we consider the weighted mean $\bar{y} = (N_1\bar{y}_1 + \cdots + N_k\bar{y}_k)/(N_1 + \cdots + N_k)$ where $\bar{y}_i$ is the mean of the n_i sampled observations in the ith sampled cluster. As this is a ratio estimator, we utilize the Taylor series linearization estimator of its repeated-sampling variance given by (Cochran (1977), p.305):

$$\hat{\mathrm{Var}}_{wor}(\bar{y}) = \frac{(1 - f_1)s_1^2 + \frac{f_1}{k}\sum_{i=1}^{k}(1 - f_{2i})N_i^2 s_{2i}^2/n_i}{k(\frac{1}{k}\sum_{i=1}^{k} N_i)^2},$$

where $s_1^2 = (k-1)^{-1}\sum N_i^2(\bar{y}_i - \bar{y})^2$, $s_{2i}^2 = (n_i - 1)^{-1}\sum(y_{ij} - \bar{y}_i)^2$, and the finite-population correction factors are $(1 - f_1) = (K - k)/K$ and $(1 - f_{2i}) = (N_i - n_i)/N_i$. This estimator requires $g(\cdot) \geq 2$.

The variance estimator setting the correction factor $1 - f_1$ equal to one corresponds to the estimator that would have been used if the sampling had actually been simple random sampling *with* replacement:

$$\hat{\mathrm{Var}}_{wr}(\bar{y}) = \frac{s_1^2}{k(\frac{1}{k}\sum_{i=1}^{k} N_i)^2}.$$

Since none of these estimators are unbiased, we consider the asymptotic case as $k, K \to \infty$ and the sampling fraction $k/K \to \gamma$ (see Appendix B). Under the

superpopulation model, we have

$$\lim_{k\to\infty} k\mathrm{Var}(\bar{y}) = \lim_{k\to\infty} kE[\hat{\mathrm{Var}}_{wr}(\bar{y})]$$
$$= \frac{1}{E_G(N)^2}\Big\{E_G[\frac{N^2\sigma^2}{g(N)}]+E_G(N^2\mu^2)+\frac{E_G(N\mu)^2E_G(N^2)}{E_G(N)^2}-2\frac{E_G(N\mu)E_G(N^2\mu)}{E_G(N)}\Big\},$$

but,

$$\lim_{k\to\infty} kE(\hat{\mathrm{Var}}_{wor}(\bar{y})) = \frac{1}{E_G(N)^2}\Big\{E_G[\frac{N^2\sigma^2}{g(N)}] + (1-\gamma)E_G(N^2\mu^2)$$
$$+(1-\gamma)\frac{E_G(N\mu)^2E_G(N^2)}{E_G(N)^2} - 2(1-\gamma)\frac{E_G(N\mu)E_G(N^2\mu)}{E_G(N)} - \gamma E_G(N\sigma^2)\Big\}.$$

Therefore, we see that ignoring the finite-population correction factors yields an asymptotically unbiased variance estimator. The asymptotic bias of $\hat{\mathrm{Var}}_{wor}(\bar{y})$ can be nonnegligible. For example, suppose the cluster sizes and cluster means are $(N_i, m_i) = (10, 1)$ or $(20, 2)$ with probability $1/2$, and $t_i^2 \equiv 1$ and $n_i \equiv 5$. Then its relative asymptotic bias is $59.44\gamma/94.44$ equaling 1%, 6%, or 16% for sampling fractions 1%, 10%, or 25% respectively.

We note in passing that the Horvitz-Thompson estimator of the mean, $\bar{y} = [(N_1\bar{y}_1 + \cdots + N_k\bar{y}_k)/k]/[(N_1 + \cdots + N_K)/K]$, is a possible estimator of the superpopulation mean when the N_i are known for all the clusters in the population. The results described above do not apply to this estimator and its usual (repeated-sampling) variance estimators. In particular, both the variance estimators that have and do not have the finite-population correction factors are asymptotically biased, with the sign of the bias depending upon the distribution G.

Case 3.2. Simple random sampling without replacement of k final units

For this case the sampling ignores the clusters in the finite population. To explain the main ideas for this case it is convenient to assume that the primary clusters in the finite population are of constant size ($N_i \equiv N_o$). Then the $\mathrm{Var}(\bar{y}) = (1/k)E_G(\sigma^2) + \{[N_o(K-1) + k(N_o-1)]/[k(N_oK-1)]\}\mathrm{Var}_G(\mu)$ and the $E[\hat{\mathrm{Var}}_{wr}(\bar{y})] = \mathrm{Var}(\bar{y}) - [(N_o-1)/(N_oK-1)]\mathrm{Var}_G(\mu)$, where $\bar{y}$ is the simple mean of the sampled units and $\hat{\mathrm{Var}}_{wr}(\bar{y})$ is the with-replacement simple-random-sampling variance estimator (2.2). We see that in general this variance estimator is an underestimate of the variability of $\bar{y}$; the without-replacement variance estimator (2.1) is even more biased. Of course, if $N_o = 1$, then we are back to case 2.1 and the with-replacement estimator is unbiased. To examine the bias in the general case, consider the relative bias:

$$\frac{E\{\hat{\mathrm{Var}}_{wr}(\bar{y})\} - \mathrm{Var}(\bar{y})}{\mathrm{Var}(\bar{y})} = \frac{-(k/K)(N_o-1)\rho}{(N_o - 1/K) + \rho(N_o-1)(k-1)/K},$$

where $\rho = \mathrm{Var}_G(\mu)/[\mathrm{Var}_G(\mu) + E_G(\sigma^2)]$ is the intraclass correlation coefficient. In our experience, when population cluster sizes are large (like geographic regions), ρ tends to be small. When cluster sizes are small (like families), the expected number of sampled units per population cluster (k/K) tends to be small; equivalently, the sampling fraction is small. In either case, the bias in ignoring the clusters when doing the sampling and estimation may not be too large. Additionally, the with-replacement variance estimator rather than the without-replacement estimator should be used.

Case 3.3. Stratified two-stage cluster sampling using simple random sampling without replacement

Let K_h be the number of primary clusters in stratum h in the finite population. At the first stage of sampling, $k_h = c_h(K_h)$ clusters are sampled from stratum h as a simple random sample without replacement, where $c_h(1) = 1$ and $c_h(t) \geq 2$ for $t \geq 2$. At the second stage of sampling, $n_{hi} = g_h(N_{hi})$ observations are sampled as a simple random sample without replacement from the (hi)th sampled cluster, where N_{hi} is the number of population units in this cluster. Let $\bar{y}_{hi}$ be the mean of these sampled observations. The weighted mean estimator of μ_{SP} is

$$\bar{y} = \sum_{h=1}^{L} \frac{K_h}{k_h} \sum_{i=1}^{k_h} N_{hi}\bar{y}_{hi} \Big/ \sum_{h=1}^{L} \frac{K_h}{k_h} \sum_{i=1}^{k_h} N_{hi}.$$

Ignoring the finite-population correction factors, the repeated-sampling variance estimator of $\bar{y}$ is (Kish (1965), p. 192):

$$\hat{\mathrm{Var}}_{wr}(\bar{y}) = \frac{\sum_{h=1}^{L} \frac{K_h^2}{k_h(k_h-1)} \sum_{i=1}^{k_h} \left[N_{hi}(\bar{y}_{hi} - \bar{y}) - \frac{1}{k_h}\sum_{j=1}^{k_h} N_{hj}(\bar{y}_{hj} - \bar{y})\right]^2}{\left(\sum_{h=1}^{L} \frac{K_h}{k_h} \sum_{i=1}^{k_h} N_{hi}\right)^2}.$$

Considering asymptotics as $K \to \infty$ and $k_h \to \infty$ (see Appendix B) with L fixed, one can show that

$$\lim_{K\to\infty} K\mathrm{Var}(\bar{y}) = \lim_{K\to\infty} KE(\hat{\mathrm{Var}}_{wr}(\bar{y})) + \Delta_{st-c},$$

where $\Delta_{st-c} = \frac{\sum_{h=1}^{L} \pi_h [E_G[N\mu|\eta=h] - \mu_{SP} E_G[N|\eta=h]]^2}{[\sum_{h=1}^{L} \pi_h E_G[N|\eta=h]]^2}$ and $\pi_h = P_G(\eta = h)$. As expected (from case 2.2), the with-replacement variance estimator asymptotically underestimates the variance; the without-replacement variance estimator is even more biased. We can asymptotically correct for the underestimation of $\hat{\mathrm{Var}}_{wr}(\bar{y})$ by using $\hat{\mathrm{Var}}_{SP}(\bar{y}) = \hat{\mathrm{Var}}_{wr}(\bar{y}) + \hat{\Delta}_{st-c}/K$, where

$$\hat{\Delta}_{st-c} = \frac{\sum_{h=1}^{L} \frac{K_h}{K} \left(\frac{1}{k_h}\sum_{i=1}^{k_h} N_{hi}\bar{y}_{hi} - \bar{y}\frac{1}{k_h}\sum_{i=1}^{k_h} N_{hi}\right)^2}{\left(\sum_{h=1}^{L} \frac{K_h}{K}\frac{1}{k_h} \sum_{i=1}^{k_h} N_{hi}\right)^2}.$$

Case 3.4. Stratified probability-proportional-to-size (pps) sampling without replacement of clusters

This is similar to case 3.3, only now we have a "size" cluster-level variable Z that can be used for differential selection probabilities. At the first stage of sampling, k_h clusters are sampled from stratum h as a pps sample without replacement. That is, the ratio of inclusion probabilities for any two clusters in stratum h is the ratio of their Z values. Cochran (1977), pp.258-270 discusses some possible ways of taking a pps without-replacement sample. We assume that the strata are formed in such a manner that it is possible to perform the stratified pps sampling. For example, if $k_h \equiv 2$, then we cannot have $K_h = 1$.

At the second stage of sampling, $n_{hi} = g_h(N_{hi}) \geq 2$ observations are sampled as a simple random sample without replacement from the ith sampled cluster from the hth stratum. Let $\bar{y}_{hi}$ be the mean of these sampled observations. The weighted mean estimator of μ_{SP} is

$$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h} N_{hi}\bar{y}_{hi}/\lambda_{hi}(\mathbf{Z}_h)}{\sum_{h=1}^{L}\sum_{i=1}^{k_h} N_{hi}/\lambda_{hi}(\mathbf{Z}_h)},$$

where $\mathbf{Z}_h = (Z_{h1}, \ldots, Z_{hK_h})$ and $\lambda_{hi}(\mathbf{Z}_h) = k_h Z_{hi}/(Z_{h1} + \cdots + Z_{hK_h})$ is the inclusion probability for the ith sampled cluster of stratum h. A with-replacement repeated-sampling pps estimator of the variance of $\bar{y}$ is given by

$$\hat{\mathrm{Var}}_{wr}(\bar{y}) = \frac{\sum_{h=1}^{L}\frac{k_h}{(k_h-1)}\sum_{i=1}^{k_h}\left[\left(\frac{N_{hi}\bar{y}_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} - \bar{y}\frac{N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right) - \frac{1}{k_h}\sum_{j=1}^{k_h}\left(\frac{N_{hj}\bar{y}_{hj}}{\lambda_{hj}(\mathbf{Z}_h)} - \bar{y}\frac{N_{hj}}{\lambda_{hj}(\mathbf{Z}_h)}\right)\right]^2}{\left(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right)^2}.$$

(Shah et al. (1993), pp.5-7, 25-26), and requires $k_h \geq 2$.

We first address the question of whether $\hat{\mathrm{Var}}_{wr}(\bar{y})$ can be used to estimate asymptotically the variance of $\bar{y}$. Typically, only a small number of clusters are sampled with pps sampling from each strata. To be realistic, therefore, we consider different asymptotics than considered in case 3.3 and let the number of strata grow with a fixed number of clusters sampled from each stratum. We can think of the stratum variable, η, determining L strata, with the asymptotics as $L \to \infty$ (Appendix B). From case 2.2, we know that the between-strata differences in the superpopulation means will not get reflected in the with-replacement variance estimator. However, we now show that even with no between-strata differences the with-replacement estimator is asymptotically biased. In particular, let $m_i \equiv 0, N_i \equiv 1$, and $k_h \equiv 2$. It can be shown that :

$$\lim_{L\to\infty} LE[\hat{\mathrm{Var}}_{wr}(\bar{y})] = \lim_{L\to\infty} E\left\{\frac{L}{K^2}\sum_{h=1}^{L}\left[\sum_{i=1}^{K_h}\frac{Y_{hi}^2}{\lambda_{hi}(\mathbf{Z}_h)} - \sum_{i=1}^{K_h}\sum_{j\neq i}^{K_h}\frac{\lambda_{hij}(\mathbf{Z}_h)}{\lambda_{hi}(\mathbf{Z}_h)\lambda_{hj}(\mathbf{Z}_h)}Y_{hi}Y_{hj}\right]\right\}$$

and

$$\lim_{L\to\infty} L\mathrm{Var}(\bar{y}) = \lim_{L\to\infty} E\Big\{\frac{L}{K^2}\sum_{h=1}^{L}\Big[\sum_{i=1}^{K_h}\frac{Y_{hi}^2}{\lambda_{hi}(\mathbf{Z}_h)} + \sum_{i=1}^{K_h}\sum_{j\neq i}^{K_h}\frac{\lambda_{hij}(\mathbf{Z}_h)}{\lambda_{hi}(\mathbf{Z}_h)\lambda_{hj}(\mathbf{Z}_h)}Y_{hi}Y_{hj}\Big]\Big\},$$

where Y_{hi} is the single unit in the ith cluster of the hth stratum, and $\lambda_{hij}(\mathbf{Z}_h)$ is the joint (second order) inclusion probability of sampling clusters i and j from stratum h (in this case Y_{hi} and Y_{hj}). The difference in sign of the terms for the cross summations of $\lim_{L\to\infty} LE[\hat{\mathrm{Var}}_{wr}(\bar{y})]$ and $\lim_{L\to\infty} L\mathrm{Var}(\bar{y})$ ensures that they will not in general be asymptotically equal.

Since the with-replacement variance estimator is not consistent for the variance of $\bar{y}$ (even with no strata effects), we pursue a different approach. Consider the decomposition: $\mathrm{Var}(\bar{y}) = E[\mathrm{Var}(\bar{y}|\text{finite population})] + \mathrm{Var}[E(\bar{y}|\text{finite population})]$, where the conditioning is on the Y values in the finite population. An estimator of the first term is provided by any usual finite-population variance estimator. For example, an analog of the Yates-Grundy estimator (Shah et al. (1993), pp.10-11) in the case of two-stage sampling, is given by

$$\hat{\mathrm{Var}}_{wor}(\bar{y}) = \frac{1}{K^2\bar{N}^2}\Big(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{i>j}^{k_h}\omega_{hij}(\mathbf{Z}_h)\Big\{\Big[\frac{N_{hi}\bar{y}_{hi}-\bar{y}N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\Big] - \Big[\frac{N_{hj}\bar{y}_{hj}-\bar{y}N_{hj}}{\lambda_{hj}(\mathbf{Z}_h)}\Big]\Big\}^2 + Ks_W^2\Big),$$

where $s_W^2 = \frac{1}{K}\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{N_{hi}^2(1-n_{hi}/N_{hi})s_{hi}^2}{n_{hi}\lambda_{hi}(\mathbf{Z}_h)}$, $\omega_{hij}(\mathbf{Z}_h) = [\lambda_{hi}(\mathbf{Z}_h)\lambda_{hj}(\mathbf{Z}_h)/\lambda_{hij}(\mathbf{Z}_h)] - 1$, $s_{hi}^2 = \sum_{j=1}^{n_{hi}}(y_{hij}-\bar{y}_{hi})^2/(n_{hi}-1)$, and $\bar{N} = \frac{1}{K}\sum_{h=1}^{L}\sum_{i=1}^{k_h}N_{hi}/\lambda_{hi}(\mathbf{Z}_h)$. It can be shown that

$$\lim_{L\to\infty} L(E[\hat{\mathrm{Var}}_{wor}(\bar{y})] - E[\mathrm{Var}(\bar{y}|\text{finite population})]) = 0.$$

If the sampling at the second stage had been a simple random sample *with replacement*, then the $(1 - n_{hi}/N_{hi})$ term would be absent in the definition of s_W^2.

To estimate $\mathrm{Var}[E(\bar{y}|\text{finite pop.})]$ requires slightly more work. The approach taken approximates $E(\bar{y}|\text{finite pop.})]$ by $\bar{Y}$ and therefore $\mathrm{Var}[E(\bar{y}|\text{finite pop.})]$ by $\mathrm{Var}(\bar{Y})$, where $\bar{Y}$ is the finite-population mean. We then obtain an estimate of $\mathrm{Var}(\bar{Y})$ from the sampled data. The steps in obtaining an estimate of $\mathrm{Var}(\bar{Y})$ are described in the next paragraph.

It is convenient to change notation temporarily and let Y_{ij} be the jth observation in the ith cluster in the finite population (regardless of stratum designation), $j = 1,\ldots,N_i$, $i = 1,\ldots,K$. In this notation,

$$\bar{Y} = \sum_{i=1}^{K}\sum_{j=1}^{N_i}Y_{ij}/\sum_{i=1}^{K}N_i$$

can be thought of as a ratio estimator. Its variance can be approximated with a Taylor series,

$$\mathrm{Var}(\bar{Y}) \doteq \frac{1}{[KE_G(N)]^2}\Big(\mathrm{Var}\Big[\sum_{i=1}^{K} N_i\bar{Y}_i\Big] + \mu_{SP}^2\mathrm{Var}\Big[\sum_{i=1}^{K} N_i\Big] - 2\mu_{SP}\mathrm{Cov}\Big[\sum_{i=1}^{K} N_i\bar{Y}_i, \sum_{i=1}^{K} N_i\Big]\Big),$$

where $\bar{Y}_i = (Y_{i1} + \cdots + Y_{iN_i})/N_i$. If we observed the data on all the individuals in the finite population, we could estimate $\mathrm{Var}(\bar{Y})$ by

$$\tilde{\mathrm{Var}}\,(\bar{Y}) = \frac{K}{K-1}\sum_{i=1}^{K}(N_i\bar{Y}_i - N_i\bar{Y})^2/\Big(\sum_{i=1}^{K} N_i\Big)^2.$$

Since we only observe the Y values on sampled individuals, $\tilde{\mathrm{Var}}\,(\bar{Y})$ is estimated from the stratified pps sample by replacing the finite population quantities with (design-based) estimators and subtracting a within-cluster variance component:

$$\hat{\mathrm{Var}}\,(\bar{Y})\frac{1}{K\bar{N}^2}\Big\{\frac{1}{K-1}\Big[\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{1}{\lambda_{hi}(\mathbf{Z}_h)}(N_{hi}\bar{y}_{hi} - N_{hi}\bar{y})^2\Big] - s_W^2\Big\},$$

where s_W^2 is defined as before. The $\hat{\mathrm{Var}}(\bar{Y})$ is the proposed estimator of $\mathrm{Var}[E(\bar{y}|$ finite pop.$)]$. It can be shown that

$$\lim_{L\to\infty} L\{\mathrm{Var}\,[E(\bar{y}|\text{finite pop.})] - E[\hat{\mathrm{Var}}\,(\bar{Y})]\} = 0.$$

The proposed estimator of $\mathrm{Var}(\bar{y})$ is $\hat{\mathrm{Var}}_{SP}(\bar{y}) = \hat{\mathrm{Var}}_{wor}(\bar{y}) + \hat{\mathrm{Var}}\,(\bar{Y})$.

A potential disadvantage of the estimator $\hat{\mathrm{Var}}_{SP}(\bar{y})$ is that since it involves $\hat{\mathrm{Var}}_{wor}(\bar{y})$, it requires knowledge of the joint inclusion probabilities. These may not be available to the analyst. In this situation, we offer the following ad hoc estimator based on $\hat{\mathrm{Var}}_{wr}(\bar{y})$: $\hat{\mathrm{Var}}_{SP-a}(\bar{y}) = \hat{\mathrm{Var}}_{wr}(\bar{y}) + \hat{\Delta}_{st-pps}$ where $\hat{\Delta}_{st-pps} = \hat{\mathrm{Var}}_b - \hat{\mathrm{Var}}_w$,

$$\hat{\mathrm{Var}}_b = \frac{\sum_{h=1}^{L} K_h\Big[\frac{1}{K_h}\sum_{i=1}^{k_h}\frac{N_{hi}(\bar{y}_{hi}-\bar{y})}{\lambda_{hi}(\mathbf{Z}_h)}\Big]^2}{\Big[\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\Big]^2},$$

and

$$\hat{\mathrm{Var}}_w = \frac{\sum_{h=1}^{L}\frac{k_h}{K_h(k_h-1)}\sum_{i=1}^{k_h}\Big(\frac{N_{hi}(\bar{y}_{hi}-\bar{y})}{\lambda_{hi}(\mathbf{Z}_h)} - \frac{1}{k_h}\sum_{j=1}^{k_h}\frac{N_{hj}(\bar{y}_{hj}-\bar{y})}{\lambda_{hj}(\mathbf{Z}_h)}\Big)^2}{\Big[\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\Big]^2}.$$

The heuristic argument justifying the estimator $\hat{\text{Var}}_{SP-a}(\bar{y})$ is given as follows. The variance of $\bar{y}$ can written as $\text{Var}(\bar{y}) = E[\text{Var}(\bar{y}|\text{obs. strat.})] + \text{Var}[E(\bar{y}|\text{obs. strat.})]$, where "obs. strat." refers to the observed strata in the finite population that are used for the sampling. This conditioning includes the values of the K_h, but not the finite population Y values. We estimate the first term of this decomposition by $\hat{\text{Var}}_{wr}(\bar{y})$ because this estimator does not account for the between-strata component of $\text{Var}(\bar{y})$, which is eliminated by the conditioning on the observed strata. An approximation to the second term can be derived as follows:

$$
\begin{aligned}
&\text{Var}[E(\bar{y}|\text{obs.strat.})] \\
&\doteq \text{Var}\Big(\frac{\sum_{h=1}^{L} E\Big(\sum_{i=1}^{k_h} \frac{N_{hi}\bar{y}_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\Big| \text{ obs.strat.}\Big)}{\sum_{h=1}^{L} E\Big(\sum_{i=1}^{k_h} \frac{N_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}|\text{obs.strat.}\Big)}\Big) = \text{Var}\Big(\frac{\sum_{h=1}^{L} K_h E_G[N\mu|\eta = h]}{\sum_{h=1}^{L} K_h E_G(N|\eta = h)}\Big) \\
&\doteq [KE_G(N)]^{-2}\text{Var}\Big\{\sum_{h=1}^{L} K_h\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}\Big\} \\
&= [KE_G(N)]^{-2}\Big(\sum_{h=1}^{L}\{K\pi_h(1-\pi_h)\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}^2\} \\
&\qquad - \sum\sum_{h\neq j} K\pi_h\pi_j\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}\{E_G[N\mu|\eta = j] \\
&\qquad\qquad -\mu_{SP}E_G[N|\eta = j)]\}\Big) \\
&= [KE_G(N)]^{-2}\Big(-\Big\{\sum_{h=1}^{L} K\pi_h\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}\Big\}^2 \\
&\qquad\qquad + \sum_{h=1}^{L} K\pi_h\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}^2\Big), \\
&= [KE_G(N)]^{-2}\sum_{h=1}^{L} K\pi_h\{E_G[N\mu|\eta = h] - \mu_{SP}E_G[N|\eta = h]\}^2,
\end{aligned}
$$

where $\pi_h = P_G[\eta = h]$. The terms on the right side of the last equality above can be estimated by $\hat{\Delta}_{st-pps}$. More refined estimators of $\text{Var}(\bar{y})$ may be possible.

We end the discussion of this case with the presentation of a simulation to demonstrate the properties of the estimators. The superpopulation consists of L' pre-strata, $L' = 50, 100$ and 200. The proportions of clusters in each pre-stratum are equal. The distribution of (N, μ, σ^2, Z) in a stratum are $(N = 5, \mu = -1 + \text{stratum effect}, \sigma^2 = 1, Z = 1)$ or $(N = 15, \mu = 1 + \text{stratum effect}, \sigma^2 = 1, Z = 2)$, each with probability $1/2$. The observations Y within a stratum and cluster are distributed as normal with mean μ and variance $\sigma^2 = 1$. The

finite population is derived as a simple random sample of $5L'$ clusters from the superpopulation. For the sampling of the finite population, the pre-strata are first numbered from 1 to L' . Then, pre-strata are pooled with their neighbor(s) to form the finite-population strata so that 2 clusters can be sampled pps from each finite-population stratum using Brewer's method (Cochran (1977), pp. 261-263): pre-strata of size 4 or more are not pooled, pre-strata of size 2 or less are always pooled, and pre-strata of size 3 are pooled depending upon the (three) values of Z. From each sampled cluster a simple random sample of 2 observations was selected. The results of the simulation are presented in Table 2 for the stratum effects being identically zero and for the hth stratum effect being $\Phi^{-1}[\{1+[\![50(h-1)/L']\!]\}/51]$ where Φ is the normal cumulative distribution function and $[\![x]\!]$ is the greatest integer less than x.

Table 2. Simulated variance of $\bar{y}$ and expectation of variance estimators using a two-stage stratified pps sample with 2 sampled primary sampling units (clusters) per pooled strata, two observations per sampled primary sampling unit and with average unpooled stratum population size of 5 (simulation size = 400,000) ; see text for details

Strata Effect $\equiv 0$			
	Number of unpooled strata		
	$L' = 50$	$L' = 100$	$L' = 200$
L'Variance$(\bar{y})$	.763	.758	.759
$L'E[\hat{\text{Var}}_{wor}(\bar{y})]$	.630	.628	.628
$L'E[\hat{\text{Var}}_{wr}(\bar{y})]$	.771	.769	.768
$L'E[\hat{\text{Var}}_{SP}(\bar{y})]$	.762	.761	.760
$L'E[\hat{\text{Var}}_{SP-a}(\bar{y})]$	.767	.767	.767
Strata Effect for stratum $h \equiv \Phi^{-1}[\{1+[\![50(h-1)/L]\!]\}/51]$			
	Number of unpooled strata		
	$L' = 50$	$L' = 100$	$L' = 200$
L'Variance$(\bar{y})$	.992	.993	.989
$L'E[\hat{\text{Var}}_{wor}(\bar{y})]$	.644	.642	.641
$L'E[\hat{\text{Var}}_{wr}(\bar{y})]$	.790	.788	.786
$L'E[\hat{\text{Var}}_{SP}(\bar{y})]$	.993	.991	.991
$L'E[\hat{\text{Var}}_{SP-a}(\bar{y})]$	.965	.965	.965

For the simulations with no strata effects, the without-replacement variance estimator is biased very low and the with-replacement variance estimator $\hat{\text{Var}}_{wr}(\bar{y})$ is biased slightly high. The superpopulation variance estimator $\hat{\text{Var}}_{SP}(\bar{y})$ appears unbiased with the approximate estimator $\hat{\text{Var}}_{SP-a}(\bar{y})$ biased slightly high. For the simulations with strata effects, even $\hat{\text{Var}}_{wr}(\bar{y})$ is biased substantially low because of its lack of incorporation of the strata effects. The

superpopulation variance estimator $\hat{\text{Var}}_{SP}(\bar{y})$ appears unbiased with increasing L' whereas the approximate estimator $\hat{\text{Var}}_{SP-a}(\bar{y})$ is biased low. We recommend using $\hat{\text{Var}}_{SP}(\bar{y})$ when the joint inclusion probabilities are known, and $\hat{\text{Var}}_{SP-a}(\bar{y})$ when they are not known.

Case 3.5. Multistage stratified sampling

At the first stage of sampling, $k_h = c_h(K_h)$ primary clusters are sampled from stratum h as a pps sample without replacement where the measure of size is Z. That is, the ratio of the inclusion probabilities for any two clusters in stratum h is the ratio of their Z values; (see case 3.4.)

At the second stage of sampling, $n_{hi} = g_h(\mathbf{Z}_{hi}, N_{hi}) \geq 2$ secondary clusters (SC's) are sampled as a pps sample with replacement (with respect to some SC-level size variable) from the ith sampled primary cluster from the hth stratum. In particular, the sample of SC's may be a simple random sample with replacement. If the same SC is sampled more than once, a completely independent sample of observations is taken from the sampled SC each time it is selected. The sampling schemes for the third to final stage of sampling can be any probability sampling scheme as long as there exists unbiased estimators (conditional on having sampled SC hij from the population) t_{hij} and d_{hij} of T_{hij}/γ_{hij} and M_{hij}/γ_{hij}, where γ_{hij} is the conditional inclusion probability for SC hij given the selection of the ith sampled primary cluster from the hth stratum. We will assume that $m_{hij} = b_{hi}(\mathbf{Z}_h, N_{hi}, M_{hij})$ observations are sampled from SC hij. Typically

$$t_{hij} = \lambda_{hi}(\mathbf{Z}_h) \sum_{l=1}^{m_{hij}} w_{hijl} y_{hijl}$$

and

$$d_{hij} = \lambda_{hi}(\mathbf{Z}_h) \sum_{l=1}^{m_{hij}} w_{hijl},$$

where w_{hijl} is the final sample weight, and $\mathbf{Z}_h$ and $\lambda_{hi}(\mathbf{Z}_h)$ are as defined in Case 3.4. The weighted mean estimator of μ_{SP} is

$$\bar{y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{k_h} \sum_{j=1}^{n_{hi}} \frac{t_{hij}}{\lambda_{hi}(\mathbf{Z}_h)}}{\sum_{h=1}^{L} \sum_{i=1}^{k_h} \sum_{j=1}^{n_{hi}} \frac{d_{hij}}{\lambda_{hi}(\mathbf{Z}_h)}}.$$

We define the following estimators $t_{hi} = t_{hi1} + \cdots + t_{hin_{hi}}$ and $d_{hi} = d_{hi1} + \cdots + d_{hin_{hi}}$. An estimator for the repeated-sampling variance of $\bar{y}$ under the assumption that the primary clusters are sampled with replacement is

$$\hat{\text{Var}}_{wr}(\bar{y}) = \frac{\sum_{h=1}^{L} \frac{k_h}{(k_h-1)} \sum_{i=1}^{k_h} \left[\left(\frac{t_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} - \bar{y} \frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} \right) - \frac{1}{k_h} \left(\sum_{j=1}^{k_h} \frac{t_{hj}}{\lambda_{hj}(\mathbf{Z}_h)} - \bar{y} \frac{d_{hj}}{\lambda_{hj}(\mathbf{Z}_h)} \right) \right]^2}{\left(\sum_{h=1}^{L} \sum_{i=1}^{k_h} \frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} \right)^2}$$

and this estimator requires that $k_h \geq 2$.

The following estimator for the variance of $\bar{y}$ under the actual without-replacement sampling design can be derived using a first-order Taylor approximation for $\bar{y}$ and Theorem 11.2 from Cochran (1977), pp.301-302

$$\hat{\text{Var}}_{wor}(\bar{y}) = \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{i>j}^{k_h}\omega_{hij}(\mathbf{Z}_h)\left(\frac{t_{hi}-\bar{y}d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} - \frac{t_{hj}-\bar{y}d_{hj}}{\lambda_{hj}(\mathbf{Z}_h)}\right)^2 + Ks_W^2}{\left(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right)^2},$$

where $s_W^2 = \frac{1}{K}\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{n_{hi}s_{hi}^2}{\lambda_{hi}(\mathbf{Z}_h)}$, $\omega_{hij}(\mathbf{Z}_h)$ is as given in case 3.4 and involves the second-order inclusion probabilities of the PSU's, and

$$s_{hi}^2 = \frac{1}{(n_{hi}-1)}\sum_{j=1}^{n_{hi}}[(t_{hij} - \bar{y}d_{hij}) - (t_{hi} - \bar{y}d_{hi})/n_{hi}]^2.$$

Let

$$\hat{\text{Var}}\,(\bar{Y}) = \frac{\frac{K}{K-1}\left[\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{1}{\lambda_{hi}(\mathbf{Z}_h)}(t_{hi} - \bar{y}d_{hi})^2\right] - Ks_W^2}{\left(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right)^2}.$$

The proposed estimator of $\text{Var}\,(\bar{y})$ is $\hat{\text{Var}}_{SP}(\bar{y}) = \hat{\text{Var}}_{wor}(\bar{y}) + \hat{\text{Var}}\,(\bar{Y})$.

An ad hoc estimator of the population variance for $\bar{y}$ that is based on $\hat{\text{Var}}_{wr}(\bar{y})$ is $\hat{\text{Var}}_{SP-a}(\bar{y}) = \hat{\text{Var}}_{wr}(\bar{y}) + \hat{\Delta}_{st-mpps}$ (*mpps* refers to multistage *pps* sampling) where $\hat{\Delta}_{st-mpps} = \hat{\text{Var}}_b - \hat{\text{Var}}_w$,

$$\hat{\text{Var}}_b = \frac{\sum_{h=1}^{L}\frac{1}{K_h'}\left[\sum_{i=1}^{k_h}\frac{(t_{hi}-\bar{y}d_{hi})}{\lambda_{hi}(\mathbf{Z}_h)}\right]^2}{\left(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right)^2} \quad \text{and}$$

$$\hat{\text{Var}}_w = \frac{\sum_{h=1}^{L}\frac{k_h}{K_h(k_h-1)}\sum_{i=1}^{k_h}\left[\left(\frac{t_{hi}}{\lambda_{hi}(\mathbf{Z}_h)} - \bar{y}\frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right) - \frac{1}{k_h}\left(\sum_{j=1}^{k_h}\frac{t_{hj}}{\lambda_{hj}(\mathbf{Z}_h)} - \bar{y}\frac{d_{hj}}{\lambda_{hj}(\mathbf{Z}_h)}\right)\right]^2}{\left(\sum_{h=1}^{L}\sum_{i=1}^{k_h}\frac{d_{hi}}{\lambda_{hi}(\mathbf{Z}_h)}\right)^2}.$$

4. An Application Using the 1987 National Health Interview Survey (NHIS)

The NHIS is a primary source of health information on the United States civilian noninstitutionalized population. Although it has been in continuous operation since 1957, the sampling design was modified in 1985 as follows (Massey et al. (1989):) The country is divided into approximately 1900 geographically defined primary sampling units (PSU's), each consisting of a county, a small group of counties, or a metropolitan statistical area. Fifty-two of the largest PSU's are sampled with certainty; these are referred to as self-representing PSU's. The

remaining PSU's are grouped into 73 strata, from each of which two PSU's are sampled without-replacement with probability-proportional-to-size. Within each sampled PSU, secondary sampling units consist of census enumeration districts, which are further subsampled leading eventually to all eligible individuals in a sampled housing unit being interviewed. We consider estimating standard errors for weighted mean estimators of family income (INCOME) and individual height in inches (HEIGHT), and the weighted proportion of females (SEX) using the data from the 1987 NHIS Cancer Control Supplement (Schoenborn and Marano (1988)).

The first line of Table 3 contains the estimated standard errors for the means of three variables using classical repeated-sampling variance estimators, which are recommended by the National Center for Health Statistics (Massey et al. (1989)). We utilize the superpopulation model of section 3 to estimate the standard errors of the superpopulation means (case 3.5) that are presented in the second line of Table 3. By doing so, we assume interest focuses on the standard errors associated with the superpopulation model rather than the repeated samping (see section 5 for further discussion). The classical variance estimators and superpopulation variance estimators differ enormously for INCOME, greatly for HEIGHT, and slightly for SEX. The large differences observed are at first surprising since the sampling fraction of PSU's is relatively small, 10% = 198/1894. However, the fifty-two self-representing PSU's (with a sampling fraction of 1.0) represent over 80% of the population, explaining the large differences. The underestimation of the design-based variances is smaller for SEX than INCOME because the variability of the PSU-level means (compared to the within-PSU variability) is greater for INCOME than SEX (data not shown).

As described previously, in the situation in which the joint inclusion probabilities of the PSU's are not available to the analyst (unlike in this survey), a with-replacement variance formula is typically used. Application of with-replacement variance formulas requires at least two sampled PSU's per stratum. To apply them to surveys in which some of the strata have only one sampled PSU (as in the present case with the self-representing PSU's), there are two techniques that are commonly used. One technique is to pair strata and their sampled PSU's so that there will be two sampled PSU's in each pseudo-stratum. The other technique is to divide the single sampled PSU into two or more pseudo-PSU's so that there will be at least two pseudo-PSU's in each stratum. In line 3 of Table 3 we present the with-replacement variance estimator with certainty PSU's paired; in line 5 we present the estimator with certainty PSU's divided into pseudo-PSU's defined by census enumeration districts. As might be expected, pairing strata (which incorporates between PSU variability) leads to estimators that are closer to the superpopulation estimators than dividing the PSU's. Line 4 contains our approximation to the superpopulation variance when the strata are paired; this

approximate variance formula does not work well when the PSU's are divided into pseudo-PSU's. For this example, the approximate formula is almost exactly the same as the formula that requires knowledge of the joint inclusion probabilities.

Table 3. Standard errors for weighted means of three variables calculated using different variance estimators (data from the 1987 National Health Interview Survey)

	INCOME ($\bar{y} = 28064$)	HEIGHT (in.) ($\bar{y} = 66.9$)	SEX (proportion female) ($\bar{y} = .53$)
Estimator			
original data:			
$\sqrt{\hat{\text{Var}}_{wor}(\bar{y})}$	193	.0328	.00402
$\sqrt{\hat{\text{Var}}_{SP}(\bar{y})}$	515	.0496	.00437
certainty PSU's paired:			
$\sqrt{\hat{\text{Var}}_{wr}(\bar{y})}$	385	.0461	.00417
$\sqrt{\hat{\text{Var}}_{SP-a}(\bar{y})}$	515	.0497	.00440
certainty PSU's divided:			
$\sqrt{\hat{\text{Var}}_{wr}(\bar{y})}$	197	.0335	.00404

5. Discussion

The variance estimators for the mean presented in the paper may be extended to other parameters of interest that can be expressed as explicit or implicit functions of means. For example, a linear regression coefficient can be expressed as a function of means of Y_i, X_iY_i, X_i^2, etc. Substitution of appropriately weighted means calculated from the sampled data can yield an estimator of the parameter. Taylor series linearization can then be used to estimate the variance of the parameter estimator by expressing the variance in terms of estimated variances of means (Binder (1983)). If interest focuses on the superpopulation parameter, then superpopulation variances for the means should be used in the Taylor series linearization. The variance estimators of this paper then apply directly. For example, suppose one is interested in a superpopulation regression coefficient in the context of the model of section 3:

$$\beta_{SP} = \frac{[E_G(N\mathcal{J}_{XY})/E_G(N\mathcal{M})] - [E_G(N\mathcal{J}_X)/E_G(N)][E_G(N\mathcal{J}_Y)/E_G(N\mathcal{M})]}{[E_G(N\mathcal{J}_{XX})/E_G(N\mathcal{M})] - [E_G(N\mathcal{J}_X)/E_G(N\mathcal{M})]^2}.$$

Here, the random vector $(N, \mathcal{M}, \mathcal{J}_{XY}, \mathcal{J}_X, \mathcal{J}_Y, \mathcal{J}_{XX})$ is at the primary cluster level and has distribution G, where recall that N is the number of secondary

clusters in a primary cluster, and $\mathcal{M}$ is the mean of the number of population values in the secondary clusters in a primary cluster. The terms in brackets in the definition of β_{SP} are superpopulation means as previously defined in section 3. Variances of their estimators will need to be estimated when estimating the variance of the estimator of β_{SP} using Taylor series linearization. If it is appropriate to consider a regression model with the sampling strata included as fixed effects, then the bias of the standard with-replacement variance estimator of a within-stratum regression coefficient may be expected to be less than the bias of the standard with-replacement variance estimator of an estimator of β_{SP}. Other possible definitions of superpopulation parameters are possible if one is willing to make stronger modeling assumptions (Pfeffermann and Holmes (1985)), and these may be more useful in some applications.

Three questions might be raised with regards to using the superpopulation variance estimators described in this paper:

(1) Why should one be interested in superpopulation inference, which is based on hypothetical constructs, when repeated-sampling inference is well-defined?
(2) Since with small sampling fractions the differences between the repeated-sampling variance estimators and superpopulation variance estimators are minor, why bother with the latter when sampling from large populations?
(3) What are the properties of the variance estimators if the superpopulation models are misspecified?

We would answer the first question with another question: Supposing you sampled the whole population of interest, would you present standard errors with your parameter estimators? If you would, then you are utilizing some type of superpopulation model, or equivalently, some distributional assumptions. We see no reason why a source of variability that is accounted for when the whole population is sampled, should be ignored when only part of the population is sampled. We believe that in most scientific applications, superpopulation inference is appropriate.

For the second question, we would agree that with a small sampling fraction of the units being used for the variance estimation, there is little benefit in the added complexity of the the superpopulation variance estimators. However, as demonstrated by the application in section 4, a small sampling fraction of final units does not imply a small sampling fraction of the PSU's which are being used for the variance estimation. Frequently, large populations are sampled with stratified multistage designs that have large sampling fractions of PSU's for at least some of the strata. In these cases, the differences between the repeated-sampling variance estimators and superpopulation variance estimators may not be small.

The third question is potentially the most troublesome. Although we believe that the superpopulation model suggested in section 3 is not very restrictive, the effects of model misspecification must always be a concern. For stratification, the discussion at the end of section 2 also applies to the model of section 3: Whatever characteristics are being used to define the sampling strata, can also be assumed to exist on the units in the superpopulation. Therefore, there are essentially no strata superpopulation-model constraints. The situation concerning clusters is different. As shown in case 3.2, not using clusters that exist in the superpopulation when sampling and estimating variances can lead to underestimation of superpopulation variances. Although the discussion of case 3.2 suggests that the underestimation may be small, let us assume for a moment that it is not small. Would this lead us to abandon superpopulation inference and use repeated-sampling inference? No, for why should one ignore the superpopulation variability due to stratification and known sampling clusters just because one is missing additional variability due to unknown superpopulation clustering? We believe that the superpopulation variance estimators described in this paper, although based on model assumptions, will yield more appropriate inferences than classical repeated-sampling variance estimators.

Acknowledgement

We thank D. Midthune for his help with the computer programming and two referees for their helpful comments.

Appendix A. Proof of Unbiasedness of $\hat{\mathrm{Var}}_{SP}(\bar{y})$ for Case 2.2

$$\begin{aligned}\mathrm{Var}(\bar{y}) &= E[\mathrm{Var}(\bar{y}|f.p.)] + \mathrm{Var}[E(\bar{y}|f.p.)] \\ &= E\Big(\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{[K_h - c_h(K_h)]}{K_h} \frac{S_h^2}{c_h(K_h)}\Big) + \mathrm{Var}(\bar{Y}),\end{aligned}$$

where *f.p.* stands for finite population, and S_h^2 is the population variance for stratum h.

$$\begin{aligned}&E[\hat{\mathrm{Var}}_{SP}(\bar{y})|f.p.] \\ &= \sum_{h=1}^{L} \frac{K_h(K_h-1)}{K(K-1)} \frac{1}{c_h(K_h)} E(s_h^2|f.p.) + \frac{1}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} E(\bar{y}_h^2|f.p.) - \frac{1}{K-1} E(\bar{y}^2|f.p.) \\ &= \sum_{h=1}^{L} \frac{K_h(K_h-1)}{K(K-1)} \frac{1}{c_h(K_h)} S_h^2 + \frac{1}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} \Big(\bar{Y}_h^2 + \frac{[K_h - c_h(K_h)]}{K_h} \frac{S_h^2}{c_h(K_h)}\Big) \\ &\quad - \frac{1}{K-1} \Big(\bar{Y}^2 + \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{[K_h - c_h(K_h)]}{K_h} \frac{S_h^2}{c_h(K_h)}\Big)\end{aligned}$$

$$= \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{[K_h - c_h(K_h)]}{K_h} \frac{S_h^2}{c_h(K_h)} + \frac{S^2}{K},$$

where S^2 is the population variance. Therefore,

$$E[\hat{\mathrm{Var}}_{SP}(\bar{y})] = E\{E[\hat{\mathrm{Var}}_{SP}(\bar{y})|f.p.]\}$$
$$= E\Big(\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{[K_h - c_h(K_h)]}{K_h} \frac{S_h^2}{c_h(K_h)} + \frac{S^2}{K}\Big) = \mathrm{Var}(\bar{y})$$

since $E(S^2/K) = \mathrm{Var}(\bar{Y})$.

Appendix B. Sufficient Conditions for Asymptotic Results

In all cases we assume there is a sequence of finite populations $\Pi_\alpha(\alpha = 1, 2, \ldots)$ that are generated by the superpopulation model described for that case. The subscript α, which indexes the finite population, is now explicitly stated. A, B and C are positive finite constants.

Case 3.1. (1) $\lim_{\alpha\to\infty} K_\alpha = \infty$
(2)$\lim_{\alpha\to\infty} k_\alpha/K_\alpha = \gamma$, where $0 < \gamma \leq 1$
(3) $2 \leq N_{\alpha i} < B$ and $|Y_{\alpha ij}| < B$

Case 3.3. (1) $\lim_{\alpha\to\infty} K_\alpha = \infty$
(2) $\lim_{t\to\infty} c_h(t)/t = \gamma_h$, where $\gamma_h > 0$ for $h = 1, \ldots, L$
(3) $2 \leq N_{\alpha hi} < B$ and $|Y_{\alpha hij}| < B$

Case 3.4. (1) $\lim_{\alpha\to\infty} K_\alpha = \infty$, $\lim_{\alpha\to\infty} L_\alpha = \infty$, $\lim_{\alpha\to\infty} K_\alpha/L_\alpha = A$
(2)$B > c_{\alpha h}(t) \geq \min(2, t)$ and $g_{\alpha h}(\cdot) \geq 2$
(3) $2 \leq N_{\alpha hi} < B$, $|Y_{\alpha hij}| < B$ and $C < Z_{\alpha hi} < B$
(4)$\lim_{\alpha\to\infty} \max_h E(K_{\alpha h}/K_\alpha) = 0$

Case 3.5. (1) $\lim_{\alpha\to\infty} K_\alpha = \infty$, $\lim_{\alpha\to\infty} L_\alpha = \infty$, $\lim_{\alpha\to\infty} K_\alpha/L_\alpha = A$
(2)$B > c_{\alpha h}(t) \geq \min(2, t)$, $B > g_{\alpha h}(\cdot, \cdot) \geq 2$, and $B > b_{\alpha hi}(\cdot, \cdot, \cdot)$
(3)$\sum_{j=1}^{N_{\alpha hi}} M_{\alpha hij} < B$, $|Y_{\alpha hij}| < B, |t_{\alpha hij}| < B$, $|d_{\alpha hij}| < B$, $|\gamma_{\alpha hij}^{-1}| < B$, and $C < Z_{\alpha hi} < B$
(4)$\lim_{\alpha\to\infty} \max_h E(K_{\alpha h}/K_\alpha) = 0$.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Internat. Statist. Rev.* **51**, 279-292.

Cassel, C., Sarndal, C. and Wretman, H. H. (1977). *Foundations of Inference in Survey Sampling.* Wiley, New York.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.* **17**, 164-177.

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

Deming, W. E. (1953). On the distinction between enumerative and analytic surveys. *J. Amer. Statist. Assoc.* **48**, 244-255.

Deming, W. E. and Stephan, F. F. (1941). On the interpretation of censuses as samples. *J. Amer. Statist. Assoc.* **36**, 45-49.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya* **87**, 117-132.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal (with discussion). *Statist. Sci.* **9**, 55-93.

Hartley, H. O. and Sielken, Jr., R. L. (1975). A 'super-population viewpoint' for finite population sampling. *Biometrics* **31**, 411-422.

Kish, L. (1965). *Survey Sampling.* Wiley, New York.

Koop, J. C. (1985). Some problems of statistical inference from sample survey data for analytic studies. *Statist.* **17**, 237-247.

Kott, P. S. (1993). Comment on Potthoff, Woodbury, and Manton. *J. Amer. Statist. Assoc.* **88**, 716.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Massey, J. T., Moore, T. F., Parsons, V. L. and Tadros, W. (1989). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, *Vital Health Statist.* 2(110).

Pfeffermann, D. and Holmes, D. J. (1985). Robustness considerations in the choice of inference for regression analysis of survey data. *J. Roy. Statist. Soc. Ser. A* **148**, 268-278.

Potthoff, R. F., Woodbury, M. A. and Manton, K. G. (1992). 'Equivalent sample size' and 'equivalent degrees of freedom' refinements for inference using survey weights under super-population models. *J. Amer. Statist. Assoc.* **87**, 383-396.

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

Shah, B. V., Folsom, R. E., LaVange, L. M., Wheeless, S. C., Boyle, K. E. and Williams, R. L. (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*, Research Triangle Institute, Research Triangle Park, NC.

Schoenborn, C. A. and Marano, M. (1988). Current estimates from the National Health Interview Survey: United States 1987. National Center for Health Statistics, *Vital Health Statist.* 10(190).

Yates, F. (1981). *Sampling Methods for Censuses and Surveys, Fourth Edition.* Charles Griffin & Company, London.

Biometric Research Branch EPN-739, National Cancer Institute, Bethesda MD 20892, U.S.A.

E-mail: korne@ctep.nci.nih.gov

E-mail: bgip@nih.gov

(Received November 1996; accepted November 1997)