VARIANCE ESTIMATION FOR SUPERPOPULATION PARAMETERS: SHOULD ONE USE WITH-REPLACEMENT ESTIMATORS?

Edward L. Korn and Barry I. Graubard, National Cancer Institute
Edward L. Korn, NCI, 6130 Executive Blvd., Bethesda Md. 20892

Key Words: Finite-population, Cluster sampling, Probability-proportional-to-size sampling

1. Introduction

Classical sampling theory concerns inferences 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 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. This advice is easy to justify; see below.

In this note we investigate variance estimation for superpopulation parameters under some more general without-replacement sampling designs. In particular, we ask whether it is appropriate to use with-replacement variance estimators that ignore finite-population correction factors for these more general designs. 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 ad hoc modification of the with-replacement estimator is given for cases in which the joint inclusion probabilities are not known to the analyst.

We restrict attention to variance estimation of the superpopulation mean in the next section, and consider other parameters in the Discussion.

2. Superpopulation models and variance estimators for the mean

For the cases that follow, the finite-population mean is always defined as the simple mean of all the observations in the realized finite population. The definition of the super-population mean depends upon the assumed superpopulation model. The superpopulation mean will not depend upon the realized finite population or the realized sample from the finite population, unlike Potthoff et al. (1992); see also Kott (1993). 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. When we refer to the "repeated sampling variance" of a statistic we mean the variance of that statistic over repeated independent samples from a fixed finite population. Derivations use standard conditioning and Taylor series arguments and are omitted.

Case 1: Simple random sampling without replacement. The finite population consists of $Y_1, \ldots, Y_K$ which are assumed to be independent, with Y_i being a realization of a random variable with mean μ_i and variance σ_i^2. The (μ_i, σ_i^2) are assumed to be independent and

identically distributed from a distribution $F(\mu, \sigma^2)$. The superpopulation mean is defined as $\mu_{SP}=E(\mu)$. Let $y_1, \ldots, y_k$ be the sampled values, and $\bar{y}$ be the sample mean. An unbiased estimator of the repeated-sampling variance of $\bar{y}$ is given by

$$\hat{var}_{wor}(\bar{y}) = \frac{(1-f)}{k} \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$

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:

$$\hat{var}_{wr}(\bar{y}) = \frac{1}{k} \frac{1}{k-1} \sum_{i=1}^{k} (y_i - \bar{y})^2$$

Using the superpopulation model, we have

$$Var(\bar{y}) = E\{\hat{var}_{wr}(\bar{y})\} = [\, E(\sigma^2) + Var(\mu) \,]/k \,,$$

and

$$E\{\hat{var}_{wor}(\bar{y}) \} = (1-f) [\, E(\sigma^2) + Var(\mu) \,] /k \,.$$

confirming the advice to ignore the finite-population correction factor for superpopulation inference; see also Fuller (1975).

Case 2: Two-stage cluster sampling using simple random sampling without replacement. For cluster i, we assume that the population values $Y_{i1}, Y_{i2}, \ldots, Y_{iN_i}$ are independent and identically distributed random variables with mean μ_i and variance σ_i^2. The population is composed of K clusters, where the (N_i, μ_i, σ_i^2) are independent and identically distributed random variables with trivariate distribution $F(N, \mu, \sigma^2)$. The superpopulation mean is defined as $\mu_{SP}= E(N\mu) / E(N)$. At the first stage of sampling, a simple random sample without replacement of k 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, or $n_i = \zeta N_i$ represents a self-weighting design.

As an estimator of μ_{SP}, we consider the weighted mean

$$\bar{y} = (\frac{1}{k} \sum_{i=1}^{k} N_i \bar{y}_i) / (\frac{1}{k} \sum_{i=1}^{k} N_i)$$

where $\bar{y}_i$ is the mean of the n_i sampled observations in the ith sampled cluster. As this is a ratio estimator, the Taylor series linearization estimator of its repeated-sampling variance is given by (Cochran 1977, p 305):

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

where $s_1^2 = \frac{1}{k-1} \sum_{i=1}^{k} N_i^2 \, (\bar{y}_i - \bar{y})^2$,

$s_{2i}^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$, and the finite-population correction factors are $(1-f_1)=(K-k)/K$ and $(1-f_{2i}) = (N_i-n_i)/N_i$. The variance estimator setting the correction factor f_1 equal to zero corresponds to the estimator that would have been used if the sampling had actually been simple random sampling with replacement:

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

Since none of these estimators are unbiased, we consider the asymptotic case as $k,K \to \infty$ and the sampling fraction $k/K \to \gamma$. Under the superpopulation model we have

$$\lim_{k\to\infty} k \, Var(\bar{y}) = \lim_{k\to\infty} k \, E (\hat{var}_{wr}(\bar{y}))$$

$$= \frac{1}{E(N)^2} [\, E(\frac{N^2\sigma^2}{g(N)}) + E(N^2\mu^2)$$

$$+ \frac{E(N\mu)^2 E(N^2)}{E(N)^2} - 2 \frac{E(N\mu) \, E(N^2\mu)}{E(N)} \,] \,,$$

but,

$$\lim_{k\to\infty} k \, E (\hat{var}_{wor}(\bar{y})) = \frac{1}{E(N)^2} [\, E(\frac{N^2\sigma^2}{g(N)})$$

$$+ (1-\gamma) \, E(N^2\mu^2) + (1-\gamma) \frac{E(N\mu)^2 E(N^2)}{E(N)^2}$$

$$-2(1-\gamma)\frac{E(N\mu)E(N^2\mu)}{E(N)} - \gamma\, E(N\sigma^2) \ ] \ .$$

Therefore, we see that ignoring the finite-population correction factors yields an asymptotically unbiased variance estimator. The asymptotic bias of $\hat{var}_{wor}(\bar{y})$, which is the estimator that would typically be used, can be non-negligible. For example, suppose the cluster sizes and cluster means are $(N_i, \mu_i) = (10, 1)$ or $(20, 2)$ with probability 1/2, and $\sigma_i^2 \equiv 1$ 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.

The stated superpopulation model for this case did not include measurement error as in Case 1. This is easily incorporated into the model (Y_{ij} has mean μ_{ij} and variance σ_{ij}^2, etc.) and the asymptotic unbiasedness of $\hat{var}_{wr}(\bar{y})$ still holds. We note also in passing that the Horvitz-Thompson estimator of the mean, $\bar{y}_{HT} = (\frac{1}{k}\sum_{i=1}^{k} N_i\bar{y}_i) \ / \ (\frac{1}{K}\sum_{i=1}^{K} N_i)$, 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, the variance estimators that both have, and do not have, the finite-population correction factors are asymptotically biased, with the sign of the bias depending upon the distribution $F(N,\mu,\sigma^2)$.

Case 3: Stratified simple random sampling without replacement. The population (and superpopulation) is composed of L disjoint strata. In stratum h, let Y_{h1}, ..., Y_{hK_h} be independent and identically distributed random variables with mean μ_h and variance σ_h^2. Unlike Case 2, we assume that these means and variances are fixed and not random since they are corresponding to strata and not clusters. The numbers of observations in the finite population falling into the different strata are assumed to be random. In particular, we assume that $(K_1, ..., K_L)$ has a multinominal distribution with sample size K and proportions $(\pi_1, ..., \pi_L)$. The superpopulation mean is defined as $\mu_{SP} = \sum \pi_h\mu_h$. From stratum h, $k_h=c_h(K_h)$ observations are sampled as a simple random sample without replacement: y_{h1}, y_{h2}, ..., 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., σ_h^2. Let

$$\bar{y} = \sum_{h=1}^{L} K_h\,\bar{y}_h \ / \ K$$

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

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

where

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

Using the superpopulation model, we have

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

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

$$\text{where } \Delta_{st} = \frac{1}{K}[\sum_{h=1}^{L}\pi_h\mu_h^2 - (\sum_{h=1}^{L}\pi_h\mu_h)^2 \ ] \ .$$

To get a feel for the magnitude of the underestimation of the variance, Table 1 presents the relative bias of (1) 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 factors, the relative biases in Table 1 would be larger by an amount equal to the sampling fraction times (1 − tabled value) .

For large sampling fractions, one may be interested in correcting the variance formula (1)

for its underestimation. This can be done by adding the following term to (1), which is an unbiased estimator of Δ_{st}:

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

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

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

With measurement error incorporated into the model, $\hat{var}_{wr}(\bar{y})$ is still biased low for $Var(\bar{y})$ by Δ_{st}, which can still be unbiasedly estimated by $\hat{\Delta}_{st}$.

Case 4: Stratified simple random sampling without replacement of clusters. This can be considered a generalization of Cases 2 and 3. The population (and superpopulation) is composed of L disjoint strata of clusters. For cluster i in stratum h, we assume that there are N_{hi} population values (Y's) that are independent and identically distributed random variables with mean μ_{hi} and variance σ_{hi}^2 . The hth stratum is composed of K_h clusters, and the $(N_{hi}, \mu_{hi}, \sigma_{hi}^2)$ are independent and identically distributed random variables with distribution function $F_h(N, \mu, \sigma^2)$. We assume that $(K_1, ..., K_L)$ has a multinomial distribution with sample size K and proportions $(\pi_1, ..., \pi_L)$. The superpopulation mean is defined by $\mu_{SP} = \sum \pi_h E_h(N\mu) \; / \; \sum \pi_h E_h(N)$, where the subscript h on the expectation denotes the expectation with respect to F_h.

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. 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. 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):

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

Considering asymptotics as $K \to \infty$ and L is fixed, one can show that

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

where $\Delta_{st\text{-}c} = \dfrac{\sum_{h=1}^{L} \pi_h [\, E_h(N\mu) - \mu_{SP} E_h(N) \,]^2}{[\, \sum_{h=1}^{L} \pi_h E_h(N) \,]^2}$

As expected (from Case 3), 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 $\hat{var}_{wr}(\bar{y})$ by using instead $\hat{var}_{SP}(\bar{y}) = \hat{var}_{wr}(\bar{y}) + \hat{\Delta}_{st\text{-}c}$, where

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

Case 5: Stratified probability-proportional-to-size (pps) sampling without replacement of clusters. This is similar to Case 4, only now we have a "size" cluster-level variable Z that can be used for differential selection probabilities. 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 4 in that now the number of strata grow with a fixed number of clusters sampled from each stratum. The population (and superpopulation) is divided into L disjoint strata of clusters. For cluster i in stratum h, we assume that there are N_{hi} population values (Y's) that are independent and identically distributed random variables with mean μ_{hi} and variance σ^2_{hi}, and we also assume the existence of a positive variable Z_{hi}. The hth stratum is composed of K_h clusters, and the $(N_{hi}, \mu_{hi}, \sigma^2_{hi}, Z_{hi})$ are independent and identically distributed random variables with distribution function $F_h(N, \mu, \sigma^2, Z)$. We assume that $(K_1, ..., K_L)$ has a multinomial distribution with sample size K and proportions $(\pi_1, ..., \pi_L)$. Since the superpopulation is the same as the number of strata grows $(L \to \infty)$, the marginal distribution $\sum_{h=1}^{L} \pi_h F_h(N, \mu, \sigma^2, Z)$ must be the same for all L, say $F_{SP}(N, \mu, \sigma^2, Z)$. (Strictly speaking the subscripts h should be Lh.) The superpopulation mean is defined by $\mu_{SP} = E_{SP}(N\mu) / E_{SP}(N)$.

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. However, we must consider the possibility that there will not be sufficient clusters in the population in a stratum to sample the required number of clusters. For example, one cannot draw a sample of $k_h=2$ clusters from a stratum that has only one cluster in it ($K_h=1$). To avoid this problem, we will pool neighboring strata so that the sampling can be done; the minimum pooling required will depend upon the particular pps sampling scheme. After pooling, let L' equal the number of strata and K'_h equal the number of clusters in the (new) hth stratum.

At the second stage of sampling, $n_{hi}=g_h(N_{hi})$ 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

$$\bar{y} = \frac{\sum_{h=1}^{L'} \sum_{i=1}^{k_h} N_{hi} \bar{y}_{hi} / \lambda_{hi}(Z_h)}{\sum_{h=1}^{L'} \sum_{i=1}^{k_h} N_{hi} / \lambda_{hi}(Z_h)}$$

where $Z_h = (Z_{h1}, \ ... \ , Z_{hK'_h})$ and $\lambda_{hi}(Z_h) = k_h Z_{hi} \Big/ \sum_{j=1}^{K'_h} Z_{hj}$ 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

$$\hat{var}_{wr}(\bar{y}) = \frac{\sum_{h=1}^{L'} \frac{k_h}{(k_h - 1)} \sum_{i=1}^{k_h} \left[d_{hi} - \frac{1}{k_h} \sum_{j=1}^{k_h} d_{hj} \right]^2}{\left(\sum_{h=1}^{L'} \sum_{i=1}^{k_h} N_{hi} / \lambda_{hi}(Z_h) \right)^2}$$

where $d_{hi} = \left(\frac{N_{hi} \bar{y}_{hi}}{\lambda_{hi}(Z_h)} - \bar{y} \frac{N_{hi}}{\lambda_{hi}(Z_h)} \right)$ (Shah et al. 1991).

We first address the question of whether $\hat{var}_{wr}(\bar{y})$ can be used to estimate asymptotically the variance of $\bar{y}$. From Case 3, we know that that between-strata differences in the superpopulation means will not get reflected in the with-replacement variance estimator. Therefore, we consider the case when the strata-specific superpopulation means of y are identically 0. For $k_h \equiv 2$, Taylor approximations yield:

$$E\left[\hat{var}_{wr}(\bar{y})\right] = \frac{1}{T^2} E\left[\sum_{h=1}^{L'} K'_h E_h \left(\frac{\mu_h^2}{\lambda_{h1}(Z_h)} \right) \right]$$

$$- \frac{1}{T^2} E\left[\sum_{h=1}^{L'} K'_h (K'_h - 1) E_h \left(\frac{\mu_{h1} \mu_{h2} \lambda_{h12}(Z_h)}{\lambda_{h1}(Z_h) \lambda_{h2}(Z_h)} \right) \right]$$

$$+ o(L^{-1})$$

$$\text{and } Var(\bar{y}) = \frac{1}{T^2} E\left[\sum_{h=1}^{L'} K'_h E_h \left(\frac{\mu_h^2}{\lambda_{h1}(Z_h)} \right) \right]$$

$$+\frac{1}{T^2}E\left[\sum_{h=1}^{L'}K'_h(K'_h-1)\,E_h\left(\frac{\mu_{h1}\mu_{h2}\,\lambda_{h12}(Z_h)}{\lambda_{h1}(Z_h)\,\lambda_{h2}(Z_h)}\right)\right]$$

$$+\ o(L^{-1})\ .$$

where $E_h[\cdot]$ is the conditional expectation with respect to the strata-specific distribution $F_h(N_h,\ \mu_h,\ \sigma_h^2,\ Z_h)$ given the pooling of the strata; $E\{\cdot\}$ on the right-hand side is the expectation over the distribution of strata poolings; $T=E[\Sigma K_h E_h(N)]$; $\lambda_{h1}(Z_h)$ and $\lambda_{h2}(Z_h)$ are the random expressions for the (first order) inclusion probabilities for the first two clusters from the hth stratum; $\lambda_{h12}(Z_h)$ is the random expression for the joint (second order) inclusion probability for the first two clusters from the hth stratum; and $\mu_{h1}$and μ_{h2} are the means for the first two clusters from the hth stratum. (Since the Z_{hi} and μ_{hi} are identically distributed, the designation of the first two clusters is arbitrary.) The difference in sign of the second terms of Var $(\bar{y})$ and $\hat{\text{var}}_{wr}$ $(\bar{y})$ insure 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 now pursue a different approach. Consider the decomposition

$$\text{Var}(\bar{y})=E[\text{Var}(\bar{y}|\ \text{finite pop., strata pooling})]$$
$$+\ \text{Var}[E\ (\bar{y}\ |\ \text{finite pop., strata pooling})],$$

where the conditioning is on the Y values in the finite population, as well as any pooling of the strata that was done. (This approach was implicitly used in the derivations for the previous cases.) 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. 1991) in the case of two-stage sampling, is given by

$$\hat{\text{var}}_{wor}(\bar{y}) = \sum_{h=1}^{L'}\sum_{i=1}^{k_h}\sum_{i>j}^{k_h}\omega_{hij}(Z_h)\ (U_{hi}-U_{hj})^2$$

$$+\sum_{h=1}^{L'}\sum_{i=1}^{k_h}\lambda_{hi}(Z_h)\ (1-n_{hi}/N_{hi})\ n_{hi}s_{hi}^2$$

where

$$\omega_{hij}(Z_h)=\left[\lambda_{hi}(Z_h)\lambda_{hj}(Z_h)\ /\ \lambda_{hij}(Z_h)\right]-1,$$

$$U_{hi}=(N_{hi}\bar{y}_{hi}-\bar{y}N_{hi})\ /\ (\hat{T}\lambda_{hi}(Z_h)),$$

$$\hat{T}=\sum_{h=1}^{L'}\sum_{i=1}^{k_h}N_{hi}\ /\ \lambda_{hi}(Z_h),$$

$$s_{hi}^2=\sum_{j=1}^{n_{hi}}(U_{hij}-\bar{U}_{hi})^2/(n_{hi}-1),$$

$$U_{hij}=(N_{hi}/n_{hi})(y_{hij}-\bar{y}N_{hi})/(\hat{T}\lambda_{hi}(Z_h)),$$

$$\bar{U}_{hi}=\sum_{j=1}^{n_{hi}}U_{hij}/n_{hi},\ \text{and}$$

$\lambda_{hij}(Z_h)$ is the joint (second order) inclusion probability of sampling PSU's i and j from stratum h. For large L, one has $E[\hat{\text{var}}_{wor}(\bar{y})]\doteq E[\text{Var}(\bar{y}|\ \text{finite pop., strata pooling})]$.

To estimate Var[$E(\bar{y}$ | finite pop., strata pooling)] requires slightly more work. For large L, E $(\bar{y}$ | finite pop., strata pooling)] $\doteq \bar{Y}$ and Var[E $(\bar{y}$ | finite pop., strata pooling)] = Var$(\bar{Y})$ + $o(L^{-1})$, where $\bar{Y}$ is the finite-population mean. We will now show how to estimate Var$(\bar{Y})$ from the sampled data. 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, ... N_i, i=1, ..., K. In this notation, $\bar{Y}=\sum_{i=1}^{K}\sum_{j=1}^{N_i}Y_{ij}\ \Big/\ \sum_{i=1}^{K}N_i$ which can be thought of as a ratio estimator. Its variance can approximated with a Taylor series,

$$\text{Var}(\bar{Y}) = \frac{1}{\left[K\,E_{SP}\,(N)\right]^2}\left(\text{Var}\left[\sum_{i=1}^{K}N_i\bar{Y}_i\right]\right.$$

$$\left.+\mu_{SP}^2\text{Var}\left[\sum_{i=1}^{K}N_i\right]-2\mu_{SP}\text{Cov}\left[\sum_{i=1}^{K}N_i\bar{Y}_i,\sum_{i=1}^{K}N_i\right]\right)$$

$$+\ o_p(K^{-1})\ ,\ \text{where}\ \bar{Y}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}Y_{ij}\ .$$

If we observed the data on all the individuals in the finite population, we could estimate Var$(\bar{Y})$ by

$$\tilde{Var}(\bar{Y}) = \frac{K}{\left[\sum_{i=1}^{K} N_i\right]^2}\left(S_Y^2 + \bar{Y}^2 S_N^2 - 2\,\bar{Y} S_{YN}\right)$$

where $S_Y^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left[N_i\bar{Y}_i - \frac{1}{K}\sum_{j=1}^{K} N_j\bar{Y}_j\right]^2$,

$S_N^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left[N_i - \frac{1}{K}\sum_{j=1}^{K} N_j\right]^2$ and

$S_{YN} = \frac{1}{K-1}\sum_{i=1}^{K}\left[N_i\bar{Y}_i - \frac{1}{K}\sum_{j=1}^{K} Y_j\right]\left[N_i - \frac{1}{K}\sum_{j=1}^{K} N_j\right]$.

Since we only observe the Y values on sampled individuals, $\tilde{Var}(\bar{Y})$ is estimated from the stratified pps sample by replacing the finite population quantities with (design-based) estimators to obtain

$$\hat{Var}(\bar{Y}) = \frac{K}{\hat{T}^2}\left(s_Y^2 - s_{YW}^2 + \bar{y}^2 s_N^2 - 2\,\bar{y}\, s_{YN}\right)$$

where $s_Y^2 = \frac{1}{K-1}\sum_{h=1}^{L'}\sum_{i=1}^{k_h}\left\{\frac{1}{\lambda_{hi}(Z_h)}\right.$

$$\left.\times\left[N_{hi}\,\bar{y}_{hi} - \frac{1}{K}\sum_{s=1}^{L'}\sum_{j=1}^{k_s} N_{sj}\,\bar{y}_{sj}\Big/\lambda_{sj}(Z_s)\right]^2\right\},$$

$s_{YW}^2 = \frac{1}{K}\sum_{h=1}^{L'}\sum_{i=1}^{k_h}\frac{N_{hi}^2(1-f_{h2i})}{\lambda_{hi}(Z_h)n_{hi}(n_{hi}-1)}\sum_{j=1}^{n_{hi}}(y_{hij}-\bar{y}_{hi})^2$,

$f_{h2i} = n_{hi}/N_{hi}$,

$s_N^2 = \frac{1}{K-1}\sum_{h=1}^{L'}\sum_{i=1}^{k_h}\left\{\frac{1}{\lambda_{hi}(Z_h)}\right.$

$$\left.\times\left[N_{hi} - \frac{1}{K}\sum_{s=1}^{L'}\sum_{j=1}^{k_s} N_{sj}\Big/\lambda_{sj}(Z_s)\right]^2\right\},$$

and $s_{YN} = \frac{1}{K-1}\sum_{h=1}^{L'}\sum_{i=1}^{k_h}\left\{\frac{1}{\lambda_{hi}(Z_h)}\right.$

$$\left.\times\left[N_{hi}\,\bar{y}_{hi} - \frac{1}{K}\sum_{s=1}^{L'}\sum_{j=1}^{k_s} N_{sj}\,\bar{y}_{sj}\Big/\lambda_{sj}(Z_s)\right]\right.$$

$$\left.\times\left[N_{hi} - \frac{1}{K}\sum_{s=1}^{L'}\sum_{j=1}^{k_s} N_{sj}\Big/\lambda_{sj}(Z_s)\right]\right\}$$

Theproposed estimator of $Var(\bar{y})$ is $\hat{var}_{SP}(\bar{y}) = \hat{var}_{wor}(\bar{y}) + \hat{Var}(\bar{Y})$ which under appropriate conditions will have the property that $\lim_{L\to\infty} L\,Var(\bar{y}) = \lim_{L\to\infty} L\,\hat{var}_{SP}(\bar{y})$.

A disadvantage of the estimator $\hat{var}_{SP}(\bar{y})$ is that since it involves $\hat{var}_{wor}(\bar{y})$, it requires knowledge of the joint inclusion probabilities. This may not be available to the analyst. In this situation, we offer the following ad hoc estimator that adds to $\hat{var}_{wr}(\bar{y})$ a between-strata variability component: $\hat{var}_{SP\text{-}a}(\bar{y}) = \hat{var}_{wr}(\bar{y}) + \hat{\Delta}_{st\text{-}pps}$, where

$\hat{\Delta}_{st\text{-}pps} =$

$$\frac{1}{\hat{T}^2}\left\{\sum_{h=1}^{L'}\frac{1}{K_h}\left[\sum_{i=1}^{k_h}\frac{\left(N_{hi}\,\bar{y}_{hi} - \bar{y}N_{hi}\right)}{\lambda_{hi}(Z_h)}\right]^2\right.$$

$$-\sum_{h=1}^{L'}\frac{k_h}{K_h(k_h-1)}\sum_{i=1}^{k_h}\left[\frac{N_{hi}\,\bar{y}_{hi}}{\lambda_{hi}(Z_h)} - \frac{1}{k_h}\sum_{j=1}^{k_h}\frac{N_{hj}\,\bar{y}_{hj}}{\lambda_{hj}(Z_h)}\right]^2$$

$$+2\bar{y}\sum_{h=1}^{L'}\frac{k_h}{K_h(k_h-1)}\sum_{i=1}^{k_h}\left[\frac{N_{hi}\,\bar{y}_{hi}}{\lambda_{hi}(Z_h)} - \frac{1}{k_h}\sum_{j=1}^{k_h}\frac{N_{hj}\,\bar{y}_{hj}}{\lambda_{hj}(Z_h)}\right]$$

$$\left.\times\left[\frac{N_{hi}}{\lambda_{hi}(Z_h)} - \frac{1}{k_h}\sum_{j=1}^{k_h}\frac{N_{hj}}{\lambda_{hj}(Z_h)}\right]\right\}$$

The idea behind this estimator is that since we know from Case 3 that the with-replacement variance estimators do not account for the between-strata variability, a better estimator would be obtained by adding an estimator of such variability. In this case, the added variability can be represented by $Var[\,E(\bar{y} \mid \text{strata pooling}, K_h\text{'s})\,]$, and $\hat{\Delta}_{st\text{-}pps}$ is an estimator of this minus its expectation under the model that there is no between-strata variability. Other ad hoc estimators are possible.

We end the discussion of Case 5 with the presentation of a simulation to demonstrate the

properties of the estimators. The superpopulation consists of L strata. The clusters are of size 1 ($N_{hi} \equiv 1$, $\sigma^2_{hi} \equiv 0$). The proportions of observations in each strata are equal ($\pi_h \equiv 1/L$). The distribution of (Y, Z) in a stratum are (Y=1+stratum effect, Z=1) or (Y= − 1+stratum effect, Z=2), each with probability 1/2. The finite population is derived as a simple random sample of 5L observations from the superpopulation. For the sampling of the finite population, the strata are first numbered from 1 to L. Then strata are pooled with their neighbor(s) so that 2 observations can be sampled pps from each finite-population stratum using Brewer's method (Cochran 1977, pp 261-263); strata of size 4 or more are not pooled, strata of size 2 or less are always pooled, and strata of size 3 are pooled depending upon the (three) values of Z. The results of the simulation are presented in Table 2 for the stratum effects being identically zero and for the hth stratum effect being $\Phi^{-1}[\ \{\ 1+\ [[\ 50\ (h\text{-}1)/L\ ]]\ \}\ /\ 51\ ]$ where Φ is the normal cumulative distribution function and [[x]] is the greatest integer less than x.

For the simulations with no strata effects, the without-replacement variance estimator is biased very low and the with-replacement variance estimator $\hat{var}_{wr}(\bar{y})$ is biased slightly high. The superpopulation variance estimator $\hat{var}_{SP}(\bar{y})$ appears unbiased with the approximate estimator $\hat{var}_{SP\text{-}a}(\bar{y})$ biased slightly high. The simulated standard deviations of the estimators $\hat{var}_{wr}(\bar{y})$, $\hat{var}_{SP}(\bar{y})$, and $\hat{var}_{SP\text{-}a}(\bar{y})$ are .073, .060, and .061, respectively for the L=200 case. For the simulations with strata effects, even $\hat{var}_{wr}(\bar{y})$ is biased substantially low because of its lack of incorporation of the strata effects. The superpopulation variance estimator $\hat{var}_{SP}(\bar{y})$ appears unbiased with increasing L whereas the approximate estimator $\hat{var}_{SP\text{-}a}(\bar{y})$ is biased slightly high. The simulated standard deviations of the estimators $\hat{var}_{wr}(\bar{y})$, $\hat{var}_{SP}(\bar{y})$, and $\hat{var}_{SP\text{-}a}(\bar{y})$ are .094, .084, and .085, respectively for the L=200 case.

We recommend for this case using $\hat{var}_{SP}(\bar{y})$ when the joint inclusion probabilities are known, and $\hat{var}_{SP\text{-}a}(\bar{y})$ when they are not known.

3. Discussion

Many parameters of interest 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 should be used in the Taylor series linearization. The results of section 2 of this paper then apply directly. For example, suppose one is interested in a superpopulation regression coefficient in the context of the clustering described above by Case 2:

$$\beta_{SP} = \frac{[E(N\mu_{XY})/E(N)] - [E(N\mu_X)/E(N)]\,[E(N\mu_Y)/E(N)]}{[E(N\mu_{XX})/E(N)] - [E(N\mu_X)/E(N)]^2}$$

Here, the cluster level vector (N, μ_{XY}, μ_X, μ_Y, μ_{XX}) is assumed to be independently and identically distributed from a multivariate distribution $F(N, \mu_{XY}, \mu_X, \mu_Y, \mu_{XX})$. The terms in brackets in the definition of β_{SP} are superpopulation means. Variances of their estimators will need to be estimated when estimating the variance of the estimator of β_{SP} using Taylor linearization. Other possible definitions of superpopulation parameters are possible if one is willing to make stronger modeling assumptions (Pfeffermann and Smith 1985), and these may be more useful in some applications.

REFERENCES

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review 51 279-292.

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

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

Cochran, W. G. (1946). Relative accuracy of sytematic and stratified random samples for a certain class of populations. Ann. Math. Statist. 17, 164-77.

Deming, W. E. & Stephan, F. F. (1941). On the interpretation of censuses as samples. J. Am. Statist. Assoc. 36, 45-9.

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

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.

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.

Pfeffermann, D. and Holmes, D. J. (1985). Robustness considerations in the choice of inference for regression analysis of survey data. J. R. Statist. Soc. 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 superpopulation 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., Barnwell, B. G., Hunt, P. N., and LaVange, L. M. (1991). SUDAAN User's Manual, Release 5.50, Research Triangle Institute, Research Triangle Park, NC.

Table 1: Relative (negative) bias of the variance estimator (1) for stratified sampling with the same sampling fractions in the different strata

	Ratio of between-strata to within-stratum variances		
Sampling Fraction	.01	.1	1
1%	<1%	1%	1%
10%	1%	5%	9%
25%	1%	7%	20%

Table 2: Simulated variance of $\bar{y}$ and expectation of variance estimators using a pps sample with 2 sampled observations per pooled strata 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($\bar{y}$)	.727	.725	.728
L E($\hat{var}_{wo}(\bar{y})$)	.524	.526	.527
L E($\hat{var}_{wr}(\bar{y})$)	.737	.738	.739
L E($\hat{var}_{SP}(\bar{y})$)	.721	.725	.726
L E($\hat{var}_{SP-a}(\bar{y})$)	.734	.737	.737

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($\bar{y}$)	.950	.951	.951
L E($\hat{var}_{wo}(\bar{y})$)	.573	.576	.577
L E($\hat{var}_{wr}(\bar{y})$)	.806	.809	.810
L E($\hat{var}_{SP}(\bar{y})$)	.944	.948	.950
L E($\hat{var}_{SP-a}(\bar{y})$)	.972	.976	.978