Limitations of Balanced Half Sampling When Strata are Grouped

Richard Valliant, Bureau of Labor Statistics
Room 4925, 2 Massachusetts Avenue NE, Washington DC 20212

Key words: replication, inconsistent variance estimator, model-based sampling, partial balancing, poststratification, two-stage cluster sampling.

1. Introduction

Balanced half-sample (*BHS*) variance estimators, and resampling estimators generally, are widely used in sample surveys because of their simplicity and flexibility. Properly applied, they can accommodate complex survey designs and complicated estimators without explicit derivations of variance formulae for different types of estimators. Thoughtless application can, however, lead to problems. This paper discusses some of the difficulties associated with *BHS* generally and with the shortcut method known as partial balancing, which can produce inconsistent variance estimators.

2. Notation and Model

The population of units is divided into H strata with stratum h containing N_h clusters. Cluster (hi) contains M_{hi} units with the total number of units in stratum h being $M_h = \sum_{i=1}^{N_h} M_{hi}$ and the total in the population being $M = \sum_{h=1}^{H} M_h$. Associated with each unit in the population is a random variable y_{hij} whose finite population total is $T = \sum_h \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} y_{hij}$. The working model is

$$E_M\left(y_{hij}\right) = \mu_h$$
$$\operatorname{cov}_M\left(y_{hij}, y_{h'i'j'}\right) = \begin{cases} \sigma_{hi}^2 & h=h', i=i', j=j' \\ \sigma_{hi}^2\rho_{hi} & h=h', i=i', j\neq j' \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

A two-stage sample is selected from each stratum consisting of $n_h = 2$ sample clusters and a subsample of m_{hi} sample units is selected within sample cluster (hi). The total number of clusters in the sample is $n = \sum_h n_h$. The set of sample clusters from stratum h is denoted by s_h and the subsample of units within sample cluster (hi) by s_{hi}.

The general estimator of the total T that we will consider in this section has the form:

$$\hat{T} = \sum_h \sum_{i\in s_h} K_{hi}\bar{y}_{hi} \qquad (2)$$

where K_{hi} is a coefficient that does not depend on the *y*'s and $\bar{y}_{hi} = \sum_{j\in s_{hi}} y_{hij}/m_{hi}$. In order for $\hat{T}$ to be model-unbiased under (1), we must have $\sum_{i\in s_h} K_{hi} = M_h$. Each K_{hi} may also depend on the particular sample selected. A number of examples of estimators that fall in the class defined by (2) were listed in Royall (1986) and Valliant (1993).

The theory here will cover the situation where H is large. If, additionally, $n/M \to 0$ and certain other population quantities are bounded as $H \to \infty$, the prediction variance $\operatorname{var}_M\left(\hat{T} - T\right)$ is asymptotically equivalent to $\operatorname{var}_M\left(\hat{T}\right)$, i.e.,

$$\operatorname{var}_M\left(\hat{T}-T\right) \approx \operatorname{var}_M\left(\hat{T}\right)$$
$$= \sum_h \mathbf{K}_h'\mathbf{V}_h\mathbf{K}_h = \sum_h \sum_{i\in s_h} K_{hi}^2 v_{hi} \qquad (3)$$

where $\mathbf{K}_h = \left(K_{h1},\ldots,K_{hn_h}\right)'$ and $\mathbf{V}_h = diag\left(v_{hi}\right)$ for $i = 1,\ldots,n_h$, and $\operatorname{var}_M\left(\bar{y}_{hi}\right) = \sigma_{hi}^2\left[1+\left(m_{hi}-1\right)\rho_{hi}\right]/m_{hi} \equiv v_{hi}$

3. A Balanced Half Sample Variance Estimator and Limits of its Applicability

Balanced half-sample (*BHS*) variance estimators, proposed by McCarthy (1969), are often used in complex surveys because of their generality and the ease with which they can be programmed. Assume that the population is stratified, as in section 2, and that a sample of $n_h = 2$ primary units is selected from each stratum. A set of *J* half-samples is defined by the indicators

$$\varsigma_{hi\alpha} = \begin{cases} 1 & \text{if cluster } hi \text{ is in half-sample } \alpha \\ 0 & \text{if not} \end{cases}$$

for *i*=1,2 and α=1,..., *J*. Based on the $\varsigma_{hi\alpha}$, define

$$\varsigma_h^{(\alpha)} = 2\varsigma_{h1\alpha} - 1$$
$$= \begin{cases} 1 & \text{if cluster } h1 \text{ is in half-sample } \alpha \\ -1 & \text{if cluster } h2 \text{ is in half-sample } \alpha \end{cases}$$

Note also that $-\varsigma_h^{(\alpha)} = 2\varsigma_{h2\alpha} - 1$. A set of half-samples is said to be in full orthogonal balance if

$$\sum_{\alpha=1}^{J} \varsigma_h^{(\alpha)} = 0, \text{ for all } h \text{ and} \qquad (4)$$
$$\sum_{\alpha=1}^{J} \varsigma_h^{(\alpha)}\varsigma_{h'}^{(\alpha)} = 0 \;(h \neq h'). \qquad (5)$$

A minimal set of half-samples satisfying (4) and (5) has $H+1 \le J \le H+4$.

Let $\hat{T}^{(\alpha)}$ be the estimator, based on half-sample α, with the same form as the full sample estimator $\hat{T}$. One of several choices of *BHS* variance estimators is

$$v_B\left(\hat{T}\right) = \sum_{\alpha=1}^{J}\left(\hat{T}^{(\alpha)} - \hat{T}\right)^2 / J.$$

3.1 Model-based Properties

Next, we can evaluate the *BHS* variance estimator and its expectation for the two-stage case. Entire clusters are assigned to half-samples, i.e., if a particular cluster is in half-sample α, then all units subsampled from that cluster are assigned to α also. The half-sample estimator of the total is defined as

$$\hat{T}^{(\alpha)} = \sum_h \left(\varsigma_{h1\alpha}K_{h1}^{(\alpha)}\bar{y}_{h1} + \varsigma_{h2\alpha}K_{h2}^{(\alpha)}\bar{y}_{h2}\right).$$

The form of the half-sample term $K_{hi}^{(\alpha)}$ is dictated by the form of $\hat{T}$ and is computed as the full sample coefficient would be if the sample size were $n_h = 1$. The α superscript is attached to $K_{hi}^{(\alpha)}$ since the value will differ from the full sample value. Although we use a superscript α on $K_{hi}^{(\alpha)}$, its value is the same for each half-sample containing unit *hi*. The difference between the half-sample and full-sample estimators is

$$\hat{T}^{(\alpha)} - \hat{T} = \sum_h \sum_{i \in s_h} \left(\varsigma_{hi\alpha} K_{hi}^{(\alpha)} - K_{hi} \right) \bar{y}_{hi} .$$

Using the definitions of $\varsigma_{hi\alpha}$ and $\varsigma_h^{(\alpha)}$, we have $\varsigma_{h1\alpha} = \left[1 + \varsigma_h^{(\alpha)}\right]/2$ and $\varsigma_{h2\alpha} = \left[1 - \varsigma_h^{(\alpha)}\right]/2$. The difference $\hat{T}^{(\alpha)} - \hat{T}$ can then be written as

$$\hat{T}^{(\alpha)} - \hat{T} = \sum_h \left\{ \left(\hat{T}_h^{(\alpha)*} - \hat{T}_h \right) + \tfrac{1}{2} \varsigma_h^{(\alpha)} \Delta_{yh}^{(\alpha)} \right\} \qquad (6)$$

where $\hat{T}_h^{(\alpha)*} = \tfrac{1}{2}\left(K_{h1}^{(\alpha)} \bar{y}_{h1} + K_{h2}^{(\alpha)} \bar{y}_{h2} \right)$, $\hat{T}_h = \sum_{i \in s_h} K_{hi} \bar{y}_{hi}$, and $\Delta_{yh}^{(\alpha)} = K_{h1}^{(\alpha)} \bar{y}_{h1} - K_{h2}^{(\alpha)} \bar{y}_{h2}$.

If $\hat{T}^{(\alpha)} - \hat{T}$ is squared out and summed over half-samples, we obtain a tidy reduction, found in McCarthy (1969) and elsewhere, *if* the K_{hi}'s and $K_{hi}^{(\alpha)}$'s have a special form, but *not* in general. In particular, suppose that

(*HS*-1) $\qquad K_{hi}^{(\alpha)} = 2K_{hi}$

holds. In that case, $\hat{T}_h^{(\alpha)*} = \hat{T}_h$, $\Delta_{yh}^{(\alpha)} = 2\Delta_{yh}$ where $\Delta_{yh} = K_{h1}\bar{y}_{h1} - K_{h2}\bar{y}_{h2}$, and

$$\hat{T}^{(\alpha)} - \hat{T} = \sum_h \varsigma_h^{(\alpha)} \Delta_{yh} . \qquad (7)$$

Squaring out (7) and summing over an orthogonal set of half-samples gives the *BHS* estimator as

$$v_B = \sum_h \Delta_{yh}^2 .$$

The expectation under model (1) is then easily calculated as

$$E_M(v_B) = \sum_h \sum_{i \in s_h} K_{hi}^2 v_{hi} + \sum_h \mu_h^2 (K_{h1} - K_{h2})^2 , \qquad (8)$$

which is the asymptotic variance in (3) plus a positive term. The positive term looks like a bias squared but is present even when $\hat{T}$ is model unbiased. If the class of estimators is further restricted so that, in addition to *HS*-1,

(*HS*-2) $\qquad K_{hi} = K_h$ for all $i \in s_h$

holds, then $\Delta_{yh} = K_h(\bar{y}_{h1} - \bar{y}_{h2})$ and $E_M(\Delta_{yh}) = 0$. With both *HS*-1 and *HS*-2 holding, v_B is approximately model unbiased.

Conditions *HS*-1 and *HS*-2 substantially limit the class of estimators for which *BHS* is appropriate as an estimator of the model variance (3). Because $\sum_{s_h} K_{hi} = M_h$ for model unbiasedness, *HS*-2 implies that $K_h = M_h/n_h = M_h/2$. In other words, the class of model-unbiased estimators for which *BHS* is appropriate consists of the singleton $\hat{T} = \sum_h M_h \sum_{s_h} \bar{y}_{hi} / n_h$.

3.2 Design-based Properties

With some sample designs v_B may have desirable design based properties when only *HS*-1 holds, despite the conditional (model) bias in (8). Define π_{hi} to be the selection probability of unit *hi* in a sample of $n_h = 2$. If $K_{hi} = M_{hi}/\pi_{hi}$, (*HS*-1) is satisfied when $K_{hi}^{(\alpha)}$ is calculated by substituting $\pi'_{hi} = \pi_{hi}/2$ for π_{hi}. In that case, $v_B = \sum_{h,s_h} \left(M_{hi}\bar{y}_{hi}/\pi_{hi} - \hat{T}_h \right)^2 / \left[n_h(n_h - 1) \right]$ and v_B is design unbiased under with-replacement sampling when $\hat{T}_h$ is design unbiased. When $K_{hi} = M_{hi}/\pi_{hi}$ and the estimator is a differentiable function of totals defined by (2), Krewski and Rao (1981) showed that v_B is design-consistent as $H \to \infty$ and the sampling of clusters is done with replacement. Condition *HS*-2 is not required for these results. When averaged over the design distribution, the second, model-related term in (8) turns into a design variance component, an example of a more general phenomenon pointed out by Smith (1994).

3.3 Examples

Some examples will show the limitations of *BHS* as an estimator of the model variance. Examples 1-2 each concern estimators of $\hat{T}$ that satisfy the condition $\sum_{i \in s_h} K_{hi} = M_h$ for unbiasedness under (1).

Example 1. *BLU* estimator: From Royall (1976) the best linear unbiased (*BLU*) predictor under (1) is $\hat{T}_{BLU} = \sum_{h,s_h} m_{hi}\bar{y}_{hi} + \sum_{h,s_h} (M_{hi} - m_{hi}) \bullet \left[w_{hi}\bar{y}_{hi} + (1 - w_{hi})\hat{\mu}_h \right] + \sum_{h,r_h} M_{hi}\hat{\mu}_h$ where r_h is the set of nonsample clusters, $w_{hi} = m_{hi}\rho_{hi}/(1 - \rho_{hi} + m_{hi}\rho_{hi})$, $\hat{\mu}_h = \sum_{s_h} u_{hi}\bar{y}_{hi}$, and $u_{hi} = q_{hi}/\sum_{s_h} q_{hi}$ with $q_{hi} = m_{hi}/\sigma_{hi}^2(1 - \rho_{hi} + m_{hi}\rho_{hi})$. The coefficient in (2) is $K_{hi} = m_{hi} + \left[M_h - m_h - \sum_{i' \in s_h} (M_{hi'} - m_{hi'}) w_{hi'} \right] u_{hi} + w_{hi}(M_{hi} - m_{hi})$ which depends on the particular units in the sample. The half-sample coefficient is simply $K_{hi}^{(\alpha)} = M_h$. Therefore, both *HS*-1 and *HS*-2 are violated.

Example 2. Horvitz-Thompson estimator when clusters are sampled with probabilities proportional to M_{hi} and an equal probability subsample is selected within each sample cluster: $\hat{T}_{HT} = \sum_h (M_h/n_h) \sum_{i \in s_h} \bar{y}_{hi}$. $K_{hi} = M_h/n_h = M_h/2$ and $K_{hi}^{(\alpha)} = M_h$, so that both *HS*-1 and *HS*-2 hold. In the special case of $\rho_{hi} = \rho_h$, $\sigma_{hi}^2 = \sigma_h^2$, $M_{hi} = \overline{M}_h$, and $m_{hi} = \overline{m}_h$, the *BLU* predictor in example 1 also reduces to $\hat{T}_{HT}$.

In both examples, the half-sample coefficients reduce to $K_{hi}^{(\alpha)} = M_h$. The same reduction occurs for the expansion estimator $\hat{T}_0 = \sum_h (M_h/m_h) \sum_{i \in s_h} m_{hi}\bar{y}_{hi}$ with $m_h = \sum_{s_h} m_{hi}$ and the ratio estimator

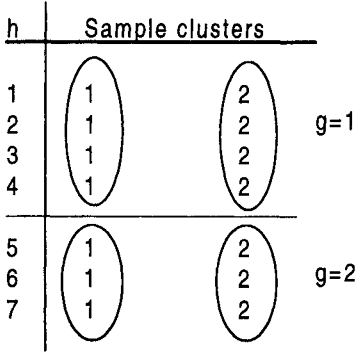
$\hat{T}_R=\sum_h\left(M_h/\sum_{i\in s_h}M_{hi}\right)\sum_{i\in s_h}M_{hi}\bar{y}_{hi}$. Thus, the half-sample method tries to estimate the variance of the *BLU* predictor, the expansion estimator, the ratio estimator, and the Horvitz-Thompson estimator all with the same set of half-sample $\hat{T}^{(\alpha)}$'s — a tactic that is obviously incorrect.

Note that standard survey design practices may minimize the effects of violating *HS*-1 and *HS*-2. If clusters are stratified based on size and the sizes M_{hi} and allocations m_{hi} are about the same within a stratum, then each of the estimators in the four estimators noted above will be approximately equal to $\hat{T}_{HT}$, the case for which *BHS* works.

4. Partial Balancing

Partial balancing is often used in order to reduce the number of half-sample estimates that must be computed for v_B. Though computationally expedient, partial balancing leads to an inconsistent variance estimator, as will be demonstrated in this section. Suppose again that $n_h=2$ and that strata are assigned to groups or superstrata. An attempt may be made to assign the same number of strata to each group, but this is not essential. In a particular group all the sample clusters numbered 1 are associated and assigned as a block to a half-sample. Sample clusters numbered 2 are similarly treated as a block. Figure 1 illustrates the grouping of strata and treatment of clusters as blocks.

Figure 1. An example of grouping strata and treating sample clusters as blocks when partial balancing is used. Circled units are assigned as a block to a half-sample.



h	Sample clusters		
1	1	2	
2	1	2	g=1
3	1	2	
4	1	2	
5	1	2	
6	1	2	g=2
7	1	2	

If there are $g=1,\ldots,\mathcal{G}$ groups of strata, then the estimator of the total can be written as $\hat{T}=\sum_{g=1}^{G}\left(\hat{T}_{g1}+\hat{T}_{g2}\right)$ where $\hat{T}_{gi}=\sum_{h\in G_g}K_{hi}\bar{y}_{hi}$, i=1,2 with G_g being the set of strata in group g. The estimator of the total based on half-sample α is

$$\hat{T}^{(\alpha)}=\sum_{g=1}^{\mathcal{G}}\left(\varsigma_{g1\alpha}\hat{T}_{g1}^{(\alpha)}+\varsigma_{g2\alpha}\hat{T}_{g2}^{(\alpha)}\right)$$

where $\varsigma_{gi\alpha}$=1 if the units numbered i in group g are in the half-sample and 0 if not, and $\hat{T}_{gi}^{(\alpha)}=\sum_{h\in G_g}K_{hi}^{(\alpha)}\bar{y}_{hi}$ with $K_{hi}^{(\alpha)}$ computed as it would be for the fully balanced case.

The difference between the grouped half-sample estimator and the full sample estimator is

$$\hat{T}^{(\alpha)}-\hat{T}=\sum_{g=1}^{\mathcal{G}}\left(\varsigma_{g1\alpha}\hat{T}_{g1}^{(\alpha)}-\hat{T}_{g1}+\varsigma_{g2\alpha}\hat{T}_{g2}^{(\alpha)}-\hat{T}_{g2}\right). \quad (9)$$

If $K_{hi}^{(\alpha)}=2K_{hi}$, i.e. *HS*-1 holds, then $\hat{T}_{gi}^{(\alpha)}=2\hat{T}_{gi}$ and

$$\hat{T}^{(\alpha)}-\hat{T}=\sum_{g=1}^{\mathcal{G}}\varsigma_g^{(\alpha)}\left(\hat{T}_{g1}-\hat{T}_{g2}\right)$$

where $\varsigma_g^{(\alpha)}=2\varsigma_{g1\alpha}-1=-\left(2\varsigma_{g2\alpha}-1\right)$. With balancing on groups, the grouped *BHS* estimator is

$$v_{GB}=\sum_g\left(\hat{T}_{g1}-\hat{T}_{g2}\right)^2.$$

The expectation of v_{GB} is easily calculated as

$$E_M\left(v_{GB}\right)=\sum_g\sum_{h\in G_g}\left(K_{h1}^2v_{h1}+K_{h2}^2v_{h2}\right)+$$

$$\sum_g\left[\sum_{h\in G_g}\mu_h\left(K_{h1}-K_{h2}\right)\right]^2, \quad (10)$$

which compares to (8) for the ungrouped case. When *HS*-2 holds, the second term in (10) is zero and the grouped *BHS* estimator is asymptotically model unbiased. Note that v_{GB} is design unbiased if only *HS*-1 holds (Wolter 1985, sec. 3.6).

Even if *HS*-1 and *HS*-2 are satisfied, v_{GB} may perform erratically when the number of groups $\mathcal{G}$ is not large. Krewski (1978), in a related case, noted the large variance of a grouped *BHS* estimator compared to the standard variance estimator in stratified simple random sampling when the stratified expansion estimator is used. Lee (1972, 1973) has studied modifications to partial balancing intended to help stabilize the variance of v_{GB}, but those procedures have somewhat limited applicability and have not become part of standard practice. Rao and Shao (1993) have also proposed a repeatedly grouped balanced half-sample (*RGBHS*) procedure that might be adapted to the partially balanced case. The *RGBHS* method applies to a case where a large number of units are selected within a stratum and then assigned at random to two groups for variance estimation.

If, as $H\to\infty$, $\mathcal{G}$ is fixed, then v_{GB} can be inconsistent in addition to being unstable. To demonstrate this, we extend an argument given by Rao and Shao (1993) and Shao (1994) for stratified single-stage sampling. Let η_g denote the number of strata assigned to group g and suppose that $\min_g\left(\eta_g\right)\to\infty$. Under standard conditions,

$$z_g=\left(\hat{T}_{g1}-\hat{T}_{g2}\right)/\sqrt{D_g}\xrightarrow{d}N(0,1)$$

where $D_g = \text{var}_M\left(\hat{T}_{g1} - \hat{T}_{g2}\right) = \sum_{h \in G_g} K_h^2 \left(v_{h1} + v_{h2}\right)$. Since $\text{var}_M\left(\hat{T} - T\right) \approx \sum_g D_g$,

$$\frac{v_{GB}}{\text{var}_M\left(\hat{T} - T\right)} \approx \sum_g \frac{D_g}{\sum_{g'} D_{g'}} z_g^2 .$$

If $D_g / \sum_{g'} D_{g'}$ converges to a constant ω_g, it follows that

$$\frac{v_{GB}}{\text{var}_M\left(\hat{T} - T\right)} \to \sum_g \omega_g \chi_g^2 , \qquad (11)$$

where χ_g^2 is a central chi-square random variable with 1 degree of freedom. In other words, rather than converging to 1 as would be the case for a consistent variance estimator, the ratio in (11) converges to a weighted sum of chi-square random variables. Note that a result similar to (11) can be obtained if $\eta_g \to \infty$ in only some of the strata. The inconsistency of v_{GB} can manifest itself by $\text{var}_M\left(v_{GB}\right)$ being large and by the length of confidence intervals being excessively variable, as verified in the simulation reported in section 6.

The occurrence in practice of this phenomenon may be more frequent than one would at first expect. In household surveys, selection of certainty clusters, i.e., selection with probability 1, is standard practice. The first-stage units in the certainties are usually geographically smaller clusters that are explicitly stratified or implicitly placed in strata through systematic sampling from an ordered list. Frequently, the first-stage sample units from a certainty are divided into two groups and v_{GB} used for variance estimation. This procedure can lead to the inconsistency described above.

5. Poststratification

Suppose that the population is divided into design strata, indexed by h, and clusters within strata as in section 2. Each unit is also a member of a class, or poststratum, denoted by c $(c = 1, \ldots, C)$. Each poststratum can cut across design strata, and the set of units in poststratum c is denoted by S_c. The total number of units in poststratum c is $M_c = \sum_h \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \delta_{hijc}$, where $\delta_{hijc} = 1$ if unit hij is in S_c and 0 if not. Assume that the poststratum sizes M_c are known. Consider the following working model

$$E_M\left(y_{hij}\right) = \mu_c \qquad (hij) \in S_c$$

$$\text{cov}_M\left(y_{hij}, y_{h'i'j'}\right) = \begin{cases} \sigma_{hic}^2 & h=h', i=i', j=j', (hij) \in S_c \\ \sigma_{hic}^2 \rho_{hic} & h=h', i=i', j \neq j', (hij) \in S_c, (h'i'j') \in S_c \\ \tau_{hicc'} & h=h', i=i', j \neq j', (hij) \in S_c, (h'i'j') \in S_{c'} \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

Let m_{hic} be the number of sample units in sample cluster hi that are part of poststratum c and $\bar{y}_{hic} = \sum_{j \in s_{hi}} y_{hij} \, \delta_{hijc} / m_{hic}$ be the sample mean of those units. The poststratified estimator is defined as

$$\hat{T}_{ps} = \sum_c \hat{R}_c \hat{T}_c$$

where $\hat{R}_c = M_c / \hat{M}_c$, $\hat{M}_c = \sum_{h,i \in s_h} K_{hic}$, and $\hat{T}_c = \sum_{h,i \in s_h} K_{hic} \bar{y}_{hic}$ with $K_{hic} = K_{hi} \, m_{hic} / m_{hi}$. A simple calculation shows that $\hat{T}_{ps}$ is unbiased under (12). Under the conditions in Valliant (1993, Appendix A.1), $\text{var}_M\left(\hat{T}_{ps} - T\right) \approx \text{var}_M\left(\hat{T}_{ps}\right)$, similar to the non-poststratified case in section 2.

Suppose that strata are grouped as in section 4 and that the *BHS* technique is used on the groups. The estimator $\hat{T}_c$ can be written as $\hat{T}_c = \sum_g \left(\hat{T}_{cg1} + \hat{T}_{cg2}\right)$ with $\hat{T}_{cgi} = \sum_{h \in G_g} K_{hic} \bar{y}_{hic}$ (i=1,2). Similarly, $\hat{M}_c = \sum_g \left(\hat{M}_{cg1} + \hat{M}_{cg2}\right)$ with $\hat{M}_{gci} = \sum_{h \in G_g} K_{hic}$. Let $K_{hic}^{(\alpha)} = K_{hi}^{(\alpha)} m_{hic} / m_{hi}$ and let $\hat{R}_c^{(\alpha)} = M_c / \hat{M}_c^{(\alpha)}$ be a half-sample poststratification ratio with $\hat{M}_c^{(\alpha)} = \sum_{g, h \in G_g} \left(\varsigma_{g1\alpha} \hat{M}_{cg1}^{(\alpha)} + \varsigma_{g2\alpha} \hat{M}_{cg2}^{(\alpha)}\right)$ and define $\hat{T}_c^{(\alpha)} = \sum_{g, h \in G_g} \left(\varsigma_{g1\alpha} \hat{T}_{cg1}^{(\alpha)} + \varsigma_{g2\alpha} \hat{T}_{cg2}^{(\alpha)}\right)$. $\hat{M}_{cgi}^{(\alpha)}$ and $\hat{T}_{cgi}^{(\alpha)}$ have the obvious definitions based on $K_{hic}^{(\alpha)}$. The half-sample poststratified estimator is $\hat{T}_{ps}^{(\alpha)} = \sum_c \hat{R}_c^{(\alpha)} \hat{T}_c^{(\alpha)}$.

When the number of strata H is large and *HS*-1 holds, $\hat{R}_c^{(\alpha)} \hat{T}_c^{(\alpha)}$ can be expanded around the full sample estimates $\hat{R}_c$ and $\hat{T}_c$ to obtain the approximation

$$\hat{R}_c^{(\alpha)} \hat{T}_c^{(\alpha)} - \hat{R}_c \hat{T}_c \cong \hat{R}_c \left[\sum_{g=1}^{\mathcal{G}} \varsigma_g^{(\alpha)} e_{gc} \right] \qquad (13)$$

where $e_{gc} = \left(\hat{T}_{cg1} - \hat{T}_{cg2}\right) - \hat{\mu}_c \left(\hat{M}_{cg1} - \hat{M}_{cg2}\right)$ with $\hat{\mu}_c = \hat{T}_c / \hat{M}_c$. After summing (13) over c, squaring, and using the orthogonality of the $\varsigma_g^{(\alpha)}$'s, the grouped *BHS* estimator is approximately

$$v_{GB} \cong \sum_{g=1}^{\mathcal{G}} \left(\sum_c \hat{R}_c e_{gc}\right)^2 = \sum_{g=1}^{\mathcal{G}} \left(\hat{\mathbf{R}}' \mathbf{e}_g\right)^2$$

with $\hat{\mathbf{R}} = \left(\hat{R}_1, \ldots, \hat{R}_C\right)'$ and $\mathbf{e}_g = \left(e_{g1}, \ldots, e_{gC}\right)'$. When (13) and *HS*-1 hold, the grouped *BHS* estimator is approximately model unbiased under (12). Note that *HS*-2 is not required for unbiasedness because the mean μ_c in model (12) does not depend on the stratum h.

Unbiasedness notwithstanding, v_{GB} is inconsistent here also. As in section 4, suppose that $\mathcal{G}$ is fixed as $H \to \infty$. Again, let η_g denote the number of strata assigned to group g and suppose that $\min_g\left(\eta_g\right) \to \infty$. Since $\hat{\mathbf{R}}' \mathbf{e}_g$ is a linear combination of random variables and each e_{gc} is a sum over a large number η_g of strata, we have, under appropriate conditions,

$$\dot{z}_g = \hat{\mathbf{R}}'\mathbf{e}_g \Big/ \sqrt{\dot{D}_g} \xrightarrow{d} N(0,1)$$

where $\dot{D}_g = \hat{\mathbf{R}}' \operatorname{var}_M(\mathbf{e}_g)\hat{\mathbf{R}}$. If $\dot{D}_g \big/ \sum_{g'} \dot{D}_{g'} \rightarrow \dot{\omega}_g$, then

$$\frac{v_{GB}}{\operatorname{var}_M\left(\hat{T}-T\right)} \rightarrow \sum_g \dot{\omega}_g \chi_g^2, \tag{14}$$

where χ_g^2 is a central chi-square random variable with 1 degree of freedom. Thus, the grouped variance estimator is also inconsistent here.

6. Simulation Results

To illustrate the problems with the grouped *BHS* variance estimator, we conducted two simulation studies. In the first, single-stage cluster sampling was used in artificial populations. In the second study, two-stage cluster samples were selected from a population derived from the U.S. Current Population Survey (CPS) and a poststratified estimator used.

For the first study, two artificial populations having $H = 40$ and $H = 160$ were generated as follows. Constant numbers of clusters per stratum and units per cluster were assigned as $N_h = 100$ and $M_{hi} \equiv \overline{M}_h = 10$. A *y* variable for each unit in each stratum was generated as $y_{hij} = \mu_h + \varepsilon_{hi} + 2\varepsilon_{hij}$ where both ε_{hi} and ε_{hij} were computed as $(x-6)/\sqrt{12}$ with x a chi-square random variable with 6 degrees of freedom. The stratum means μ_h were multiples of 10, assigned in blocks of 20 — $\mu_h = 10$ for the first 20 strata, $\mu_h = 20$ for the next 20 strata, $\mu_h = 30$ for the next 20 strata (for $H = 160$), and so on. The population with $H = 40$ had a total of $M = 40{,}000$ units while the $H = 160$ population had 160,000 units. In each stratum a sample of $n_h = 2$ was selected by simple random sampling without replacement and both sample clusters were completely enumerated. The estimator of the total used was $\hat{T} = \sum_h M_h \bar{\bar{y}}_{hs}$ with $\bar{\bar{y}}_{hs} = \sum_{s_h} \bar{y}_{hi}/n_h$. $\hat{T}$ is unbiased with respect to both the model and the stratified simple random sampling design. When the sampling fraction of clusters is small in each stratum (2/100 here), a model-unbiased and design-unbiased estimator of variance is

$$v_B = \sum_h M_h^2 \left(\bar{y}_{h1} - \bar{y}_{h2}\right)^2 / 4,$$

which also equals the *BHS* estimator when a set of half-samples in full orthogonal balance is used.

For both artificial populations v_{GB} was computed using $\mathcal{G} = 20$ groups and a set of 24 half-samples in full orthogonal balance. When $H = 40$, strata were paired to form the groups. Strata 1 and 2 were paired, strata 3 and 4 were paired and so on. When $H = 160$, strata 1-8 were grouped, strata 9-16, and so on. Note that this type of purposive, as opposed to random, grouping reflects what is typically done in practice.

The second study used a population of 10,841 persons included in the September 1988 CPS. The *y* variable was weekly wages for each person. The study population contained 2,826 geographic clusters, each composed of about 4 neighboring households. Eight poststrata were formed based on age, race, and sex (Valliant 1993). A two-stage sample design was used with clusters as first-stage units and persons as second-stage units. Two sets of 1,000 samples were selected with 100 sample clusters in the first set and 200 sample clusters in the second. In both sets, clusters were selected systematically with probabilities proportional to the number of persons in each cluster. Strata were created in both cases to have about the same total number of households, and $n_h = 2$ sample clusters selected in each stratum. In each sample cluster, a simple random sample of 4 persons was selected without replacement in clusters with $M_{hi} > 4$; otherwise, the cluster was enumerated completely.

From each sample from the CPS population, the poststratified estimate $\hat{T}_{ps}$, the *BHS* variance estimator v_B based on a set of half-samples in full orthogonal balance, and the grouped *BHS* estimator were calculated. The poststratified estimate $\hat{T}_{ps}$ used $K_{hi} = M_h / n_h$ so that *HS*-1 and *HS*-2 were satisfied. For both sample sizes (*n*=100 and *n*=200), 25 groups of strata were formed in order to compute v_{GB}. For both v_B and v_{GB}, the half-sample totals $\hat{T}_c^{(\alpha)}$ incorporated the factor $\sqrt{1 - n_h/N_h}$ to approximately reflect the effect of a non-negligible *fpc*.

Table 1 summarizes results on square root mean square errors (*rmse*'s) and standard error estimates across 1,000 samples from each of the populations. As the ratios, $\overline{v^{1/2}}/rmse$, of average root variance estimate to *rmse* show, neither the grouped *BHS* estimator nor the fully balanced choices have any serious biases in either the artificial or CPS populations.

Table 1. Empirical root mean square errors (*mse*) of estimators of totals and ratios of average standard error estimates to the *rmse* in 1,000 samples.

Population	*rmse* (000s)	$\overline{v_B^{1/2}}/rmse$	$\overline{v_{GB}^{1/2}}/rmse$
Artificial $\hat{T}_0$			
$H = 40$	5.2	1.002	.997
$H = 160$	9.9	1.051	1.040
CPS $\hat{T}_{ps}$			
$H = 50$	133.0	1.055	1.049
$H = 100$	97.4	.977	1.022

Ninety percent, 95%, and 99% confidence intervals (CI's) based on either v_B or v_{GB} covered at approximately the desired rates, but more interesting properties of CI's are given in Table 2. That table lists the averages of the half-widths of 95% CI's, i.e. the average over the samples of $1.96\sqrt{v}$ for each variance estimator v. The table also shows the variances of those half-widths. Although, for a given simulation, the average length is about the same for both variance estimators, the variances of the half-widths are vastly different. In the (artificial/*H*=40) case, the variance of the v_{GB} half-widths is 1.9 times the variance of the v_B half-widths. In the (artificial/*H*=160) case, the ratio is 6.2. The ratios of variances for the (CPS/*H*=50) and (CPS/*H*=100) cases are 2.1 and 4.4.

Table 2. Empirical results for average half-width length and variance of half-width length for 95% confidence intervals over 1,000 samples.

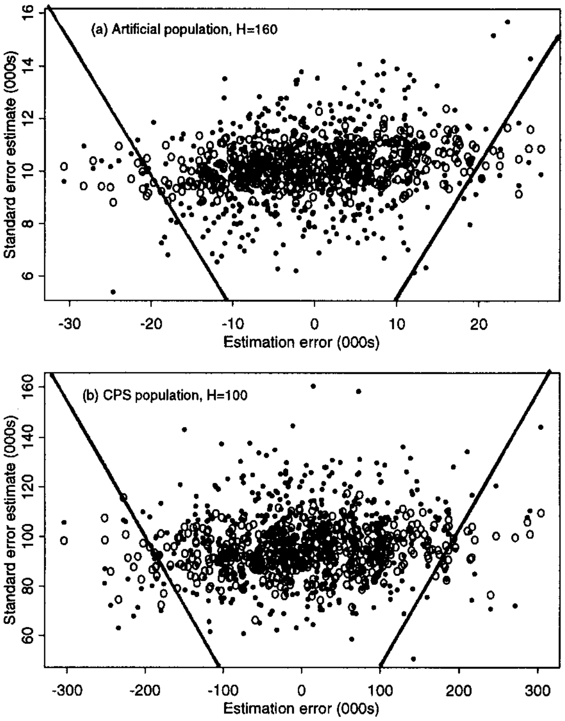
Population	Mean half width (000s)		Var of half-width (000,000s)		Ratio of half-width vars (v_{GB}/v_B)
Artificial	v_B	v_{GB}	v_B	v_{GB}	
$H=40$	10.2	10.2	1.6	3.0	1.9
$H=160$	20.3	20.1	1.6	10.1	6.2
CPS					
$H=50$	275.1	273.6	963.6	2,016.0	2.1
$H=100$	186.5	195.1	234.7	1,035.4	4.4

The relative instability of v_{GB} and its effect on confidence interval coverage and length is further illustrated by Figure 2. The standard error estimate $\sqrt{v}$ (v = v_B or v_{GB}) for each sample is plotted versus the estimation error $\hat{T}-T$ for 500 of the samples for (Artificial/*H*=160) and (CPS/*H*=100). Reference lines are drawn at $\sqrt{v}=|\hat{T}-T|/1.96$. Points that fall between the two lines correspond to samples where the 95% confidence interval covered the true value. Points outside the reference lines are samples where the confidence intervals did not cover. Circles denote v_B and dots v_{GB}. In both panels v_B has a more narrow range for almost all values of $\hat{T}-T$ than does v_{GB}. The width of confidence intervals based on v_{GB} is erratic in the region where intervals cover *T*. Near $\hat{T}-T=0$ in (CPS/*H*=100), for example, $\sqrt{v_{GB}}$ ranges from about 60 to 160 (in thousands), but the range of $\sqrt{v_B}$ is about 75 to 120.

7. Conclusion

Though balanced half-sampling can be a flexible and powerful tool in complex sample surveys, the shortcut method of partial balancing should be avoided unless a large number of groups can be formed. The grouped *BHS* variance estimator is at best unstable compared to a fully balanced estimator and at worst inconsistent. Continuing surveys that use partial balancing are likely to observe erratic point estimates of variance over time that do not accurately reflect the precision of estimated means and totals.

Figure 2. Standard error estimates ($\sqrt{v_B}$ and $\sqrt{v_{GB}}$) plotted versus estimation errors $(\hat{T}-T)$ for 500 samples from the artficial population (*H*=160) and the CPS population (*H*=100). o = $\sqrt{v_B}$; • = $\sqrt{v_{GB}}$



ACKNOWLEDGMENTS

The author thanks A. Dorfman for his careful review. Any opinions expressed here are those of the author and do not constitute policy of the Bureau of Labor Statistics.

REFERENCES

Krewski, D. (1978). On the stability of some replication variance estimators in the linear case. *Journal of Statistical Planning and Inference*, **2**, 45-51.

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

Lee, K.-H. (1972). Partially balanced designs for half sample replication method of variance estimation. *Journal of the American Statistical Association*, **67**, 324-334.

Lee, K.-H. (1973). Using partially balanced designs for the half sample replication method of variance estimation. *Journal of the American Statistical Association*, **68**, 612-614.

McCarthy, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, **37**, 239-264.

Rao, J.N.K., and Shao, J. (1993). On balanced half-sample variance estimation in stratified sampling. Carleton University preprint.

Royall, R. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657-664.

Shao, J. (1994). Resampling methods in sample surveys. *Statistics*, to appear.

Smith, T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation?" *International Statistical Review*, **62**, 3-34.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, **88**, 89-96.

Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag: New York.