BRR versus Inclusion-Probability Formulas for Variances of Nonresponse Adjusted Survey Estimates

Eric V. Slud * Yves Thibaudeau †

Abstract
This paper compares two methods of calculating estimated theoretical variances for nonresponse adjusted estimates of survey totals. Both methods employ a possibly misspecified parametric model for the nonresponse probabilities. The first method is based on a formula of Särndal and Lündstrom (2005) available when joint survey inclusion probabilities are known and when response probabilities are modelled in terms of survey variables through a calibration model. The second method is based on balanced repeated replicates (BRR). Both methods are compared with the correct design-based variance in a superpopulation framework with independent random response indicators. Numerical calculations and confirmatory simulation results are given for a split-PSU design, with simple random sampling within half-PSU's and with ratio estimators in adjustment cells used to adjust survey weights for nonresponse. The accuracy of variance estimators is exhibited in relation to the balance across half-PSU's of the intersections of true and working adjustment cell frequencies in the population.

Key Words: balanced replication, misspecified model, bias, pseudo-randomization, superpopulation, inclusion probability.

1. Introduction & Notation

Consider a sample survey with a frame $\mathcal{U}$ from which a probability sample $\mathcal{S}$ is drawn according to a plan with known single and joint inclusion probabilities π_i, π_{ij}, for $i, j \in \mathcal{U}$. Assume that the total $Y = t_y = \sum_{i\in\mathcal{U}} y_i$ of a scalar attribute is of primary interest, and that $(y_i, x_i, i \in \mathcal{S})$ is observable, i.e., the sample data includes the auxiliary p-dimensional vector x_i. This design-based setting corresponds to the *InfoS* sampling framework of Särndal and Lündstrom (2005), where auxiliary sample data are available but not population-level data.

Assume also that each sampled individual in the survey decides independently whether or not to respond (the 'quasi-randomization' model as in Oh and Scheuren 1983.). Denote by r_i for all $i \in \mathcal{U}$ the indicator which is 1 if the i'th individual *would* have responded if sampled, and assume that these random variables are independent of each other and of the sample selection mechanism. The observable data are now taken to be $(y_i\, r_i,\, r_i,\, x_i,\, i \in \mathcal{S})$. The probabilities with which individual units respond

$$P(r_i = 1) \;=\; Er_i \;\equiv\; 1/\phi_i$$

must be estimated in order to adjust weights for nonresponse, and this is typically done either by ratio-adjustment and raking (Oh and Scheuren 1983) or by using a working generalized-linear parametric model

$$\phi_i \;=\; g(\lambda' x_i) \quad , \qquad i \in \mathcal{U} \qquad , \qquad g \quad \text{known}$$

where λ is a p-vector parameter to be estimated from sample data as the solution $\hat\lambda$ to an estimating equation

$$\sum_{i\in\mathcal{S}} (1/\pi_i)\, x_i\, (r_i\, g(\lambda'\, x_i) \;-\; 1) \;=\; \mathbf{0}$$

after which the total Y is estimated through the nonresponse-adjusted weighted total

$$\hat Y_g \;=\; \sum_{i\in\mathcal{S}} (y_i/\pi_i)\; r_i\, g(\hat\lambda'\, x_i)$$

The most important example of a working nonresponse model and its predictors is

$$\phi_i \equiv (Er_i)^{-1} \;=\; g(\lambda' x_i) \;=\; 1 + \lambda'\, x_i \quad , \qquad x_i \;=\; (I_{[i\in C_1]},\, I_{[i\in C_2]},\, \ldots,\, I_{[i\in C_M]}) \;\in\; \{0,1\}^M \tag{1}$$

*Census Bureau, SRD, 4600 Silver Hill Rd Rm 5K004, Washington DC 20233-9100, & Math. Dept., Univ. of Maryland, College Park
†Census Bureau, Statist. Res. Div., 4600 Silver Hill Road Rm 5K105, Washington DC 20233-9100

where the cells $C_1, \ldots, C_M$ partition the frame population $\mathcal{U}$. For simplicity, we restrict attention throughout this paper to the ratio-adjustment calibration model (1) and its resulting weight-adjusted estimator of the total Y,

$$\hat{Y} \;=\; \sum_{i\in\mathcal{S}} \frac{y_i}{\pi_i}\, r_i\, (1+\hat{\lambda}' x_i) \;=\; \sum_{m=1}^{M} \hat{c}_m \sum_{i\in\mathcal{S}\cap C_m} \frac{y_i}{\pi_i}\, r_i \quad , \qquad \hat{c}_m \;=\; \sum_{a\in\mathcal{S}\cap C_m} (1/\pi_a) \Big/ \sum_{a\in\mathcal{S}\cap C_m} r_a/\pi_a \tag{2}$$

where the calibrated adjustment factors estimators $\hat{c}_m = 1+\hat{\lambda}' x_i$ are determined by (2) for $i \in C_m$, $1 \le m \le M$. From now on, we indicate the cell-index m for $i \in \mathcal{U}$ by $m(i)$, that is, $i \in C_m$ if and only if $m = m(i)$.

This paper is about the estimation of variance of the weight-adjusted total estimator $\hat{Y}$, in settings where the parametric working model is possibly incorrect. One source of variance formulas is Taylor linearization involving weight-adjusted inclusion probabilities, as given by Särndal and Lündstrom (2005) and extended in Section 2 below. Apart from these, the most important technique of variance estimation in general use is the class of Balanced Repeated Replicate (BRR) methods (Wolter 1985, Brick, Morganstein and Valliant 2004). These are methods which rely on split-sample symmetries in the sample design to produce unbiased or nearly unbiased variance estimates depending only on single inclusion probabilities and a finite set of multiplicative replicate factors (Kish and Frankel 1970, Fay 1984, 1989). In the simplest case, the survey can be idealized to have PSU's indexed by $k = 1, \ldots, K$, which are split into balanced halves $\mathcal{U}_{kH}$ indexed by $H = 1, 2$, so that the samples $(y_i,\, x_i,\, i \in \mathcal{S}(k,H))$ drawn from the two halves can be regarded as approximately independent and identically distributed. (This discussion is more generally applicable to the case where PSU's are sampled through a more complicated structure involving strata within which PSU's are nested.) Assume in what follows that the sample inclusion probabilities π_i for i in PSU k and half-PSU H have the form $\pi_i = \pi^*_k\, \pi_{i|k} \;=\; \pi_{i|k}/w_k$ not depending on H.

The BRR methods we consider are defined (Fay 1984, 1989) in terms of multiplicative weight factors $(f_{i,t},\, i \in \mathcal{S},\, t = 1, \ldots, R)$ which satisfy $R^{-1}\sum_{t=1}^{R} f_{i,t} = 1$ together with certain orthogonality relations. (In practice, the average-replicate equality and the orthogonality hold only approximately, because the numbers of replicates are reduced below the numbers of PSU's to reduce computation times and storage.) The replicated estimators are

$$\hat{Y}^{(t)} \;\equiv\; \sum_{i\in\mathcal{S}} \hat{c}^{(t)}_{m(i)}\, \frac{y_i}{\pi_i}\, r_i\, f_{i,t} \;=\; \sum_{m=1}^{M} \hat{c}^{(t)}_m \sum_{i\in\mathcal{S}\cap C_m} \frac{y_i}{\pi_i}\, r_i\, f_{i,t} \quad , \qquad \hat{c}^{(t)}_m \;=\; \frac{\sum_{a\in\mathcal{S}\cap C_m} f_{a,t}/\pi_a}{\sum_{a\in\mathcal{S}\cap C_m} f_{i,t}\, r_a/\pi_a} \tag{3}$$

and there is a quadratic form $Q(\{y_i\}_{i\in\mathcal{S}})$ in the sampled attributes (Fay 1984) such that the variance of the estimator $\hat{Y}$ given in (2) can be estimated, when the number R of replicates is taken large enough, by

$$\hat{V}_{BRR}(\hat{Y}) \;=\; \frac{4}{R}\sum_{t=1}^{R} (\hat{Y}^{(t)} - \hat{Y})^2 \;=\; Q(\{y_i\}_{i\in\mathcal{S}}) \tag{4}$$

The quadratic form Q varies – and must be justified as an unbiased or nearly unbiased estimator – with the application, as does the required number of replicates which is usually a small multiple of the number of PSU's. But in the balanced half-sample setting described above, it is shown by Slud and Thibaudeau (2008) that the top-order term of this estimator $\hat{V}_{BRR}(\hat{Y})$ is given in large data-samples by the quadratic form

$$\sum_{k=1}^{K} \Big(\sum_{i\in\mathcal{S}(k,1)} \frac{1}{\pi_i}\, (\hat{\beta}' x_i + r_i(1+\hat{\lambda}' x_i)\, \hat{e}_i) \;-\; \sum_{i\in\mathcal{S}(k,2)} \frac{1}{\pi_i}\, (\hat{\beta}' x_i + r_i(1+\hat{\lambda}' x_i)\, \hat{e}_i) \Big)^2 \tag{5}$$

where $\hat{\beta}$, $\hat{e}_i$ are as in formulas (7) and (8) below.

This paper is organized as follows. In the next Section, we provide theoretical formulas and estimators for variances of survey estimators weight-adjusted for nonresponse using adjustment cells, in terms of the actual and estimated response probabilities. Section 3 introduces a general framework within which to study the large-sample behavior of variance formulas and BRR variance estimators under cell-based adjustments even when the working cells may not be the correct ones. Section 4 gives numerical comparisons under various realistic scenarios and simulations which confirm the theoretical formulas. In Section 5 we provide a Summary and Discussion of results.

2. Linearization-Based Variance Formulas

We begin with linearization-based formulas and estimators for variance of the calibration survey estimator (2) based on the assumed-known joint inclusion probabilities π_{ij}. The special form of inclusion probabilities assumed above is maintained here, reflecting sampling that is independent across PSU's. The assumptions needed to prove the

formulas (*cf.* Slud and Thibaudeau 2008, Appendix) can take many forms, related to the design consistency of survey-weighted estimators (Särndal et al. 1992). For example, the formulas are valid as n, N both get large, if for some constant $\alpha > 0$ not depending on n, N

$$\max_i \phi_i \le \alpha^{-1} \quad , \quad \min_i \pi_i / \max_i \pi_i \ge \alpha \quad , \quad (\sum_i |y_i|^3)^{2/3} / (\sum_i y_i^2) \le \alpha^{-1} \quad , \quad N^{-1} \sum_{i \in \mathcal{U}} (y_i - t_y/N)^2 \ge \alpha \tag{6}$$

Because calibration and regression are closely related, we follow Särndal and Lündstrom (2005) in introducing regression coefficients, estimators and residuals. Define M-vectors β^0, $\hat\beta$ through their m'th components

$$\beta^0_m = \sum_{i \in C_m} (y_i/\phi_i) \Big/ \sum_{i \in C_m} (1/\phi_i) \quad , \qquad \hat\beta_m = \sum_{i \in \mathcal{S} \cap C_m} (y_i\, r_i/\pi_i) \Big/ \sum_{i \in \mathcal{S} \cap C_m} (r_i/\pi_i) \tag{7}$$

as m ranges from $1, \ldots, M$, and define corresponding residuals for all $i \in C_m$ by

$$e^0_i = y_i - \beta^0_m \quad , \qquad \hat e_i = y_i - \hat\beta_m \quad , \qquad e_i = y_i - \sum_{j \in C_m} y_j/|C_m| \tag{8}$$

The notation $A \approx B$ for (random) expressions A, B indicates that $A - B$ is of smaller order than A in probability as N and n become large. (That is, $A \approx B$ means that $(A-B)/|A|$ tends to 0 in probability.)

Proposition 1 (Särndal and Lündstrom 2005) *Under the assumptions listed above, the idealized estimator obtained by replacing $\hat c_{m(i)}$ in formula (2) for $\hat Y$ with ϕ_i has variance*

$$\approx \sum_{i,j \in \mathcal{U}} \Big(\frac{\pi_{ij}}{\pi_i\, \pi_j} - 1 \Big) y_i\, y_j + \sum_{i \in \mathcal{U}} (\phi_i - 1)\, (e_i^2/\pi_i) \tag{9}$$

The Proposition suggests, following Särndal and Lündstrom (2005), the simplified variance estimator

$$\hat V_{SL}(\hat Y) = \sum_{i,j \in \mathcal{S}} \Big(\frac{\pi_{ij}}{\pi_i\, \pi_j} - 1 \Big) \frac{y_i\, y_j}{\pi_{ij}} + \sum_{m=1}^{M} (\hat c_m - 1) \sum_{i \in \mathcal{S} \cap C_m} (\hat e_i/\pi_i)^2 \tag{10}$$

This 'estimator' would be applicable only if y_i for all $i \in \mathcal{S}$ were observable. Since only the attribute values y_i for $i \in \mathcal{S}$ and $r_i = 1$ are observable, the usable form of this estimator is:

$$\sum_{i,j \in \mathcal{S}} \Big(\frac{\pi_{ij}}{\pi_i\, \pi_j} - 1 \Big) \frac{r_i\, r_j\, y_i\, y_j}{\pi_{ij}}\, \hat c_{m(i)}\, (I_{[i=j]} + I_{[i \ne j]} \hat c_{m(j)}) + \sum_{i \in \mathcal{S}} (\hat c_{m(i)} - 1)\, \hat c_{m(i)}\, r_i\, (\hat e_i/\pi_i)^2 \tag{11}$$

Särndal and Lündstrom (2005) propose an estimator like (11) for $\mathrm{Var}(\hat Y)$. A more detailed expression for $\mathrm{Var}(\hat Y)$ derived by Taylor linearization shows how the validity of the approximation by (11) depends on the correctness of the working model $\phi_i = 1 + x_i' \lambda_0$, where λ_0 is the large-sample limit of $\hat\lambda$, with m'th component given by

$$(\lambda_0)_m \equiv c_m - 1 = \sum_{i \in C_m} (1 - \phi_i^{-1}) \Big/ \sum_{i \in C_m} \phi_i^{-1} \tag{12}$$

Proposition 2 (Slud and Thibaudeau 2008) *Under the same set of assumptions as in Prop. 1,*

$$Var(\hat Y) \approx \sum_{i,j \in \mathcal{U}} \Big(\frac{\pi_{ij}}{\pi_i\, \pi_j} - 1 \Big) \Big\{ (x_i' \beta^0 + \frac{e^0_i}{\phi_i}\, c_{m(i)})\, (x_j' \beta^0 + \frac{e^0_j}{\phi_j}\, c_{m(j)}) \Big\} + \sum_{i \in \mathcal{U}} \frac{(c_{m(i)}\, e^0_i)^2}{\pi_i\, \phi_i} (1 - \phi_i^{-1}) \tag{13}$$

Formula (13) is definitely more complex than (9). Yet, if $(E r_i)^{-1} = \phi_i = 1 + \lambda_0' x_i$ for all i, then there is considerable cancellation, and the approximate variances of $\hat Y$, (9) and (13), are the same if the residuals e_i appearing in (9) and e^0_i coincide, which happens if the y_i attributes are *iid*, independent of the calibration variables x_i, with mean μ and variance σ^2.

The variance expression in Prop. 2 is theoretical, and contains the unknowns ϕ_i, which can be estimated in practice only if there is a valid parametric model $\phi_i = g(x_i, \gamma)$ as assumed in Kim and Kim (2007). Validation of such a model is generally not possible without additional followup on nonresponders. Nevertheless, if a more detailed logistic-regression or adjustment-cell model were fitted to provide a more refined estimation of response probabilities, then this formula based on modelled and estimated values $\hat\phi_i = g(x_i, \hat\gamma)$, with $\hat\gamma$ derived from a sample-based estimating equation, could be used to assess the quality of the approximation $\mathrm{Var}(\hat Y) \approx \mathrm{Var}(\hat Y_G)$ proposed by Särndal and Lündstrom (2005). An estimator based on this idea is

$$\sum_{i \in \mathcal{S}} \frac{r_i}{\pi_i^2}\, \hat c^2_{m(i)}\, \hat e_i^2\, (1 - \hat\phi_i^{-1}) + \sum_{i,j \in \mathcal{S}} \Big(\frac{\pi_{ij}}{\pi_i\, \pi_j} - 1 \Big) \frac{r_i\, r_j}{\pi_{ij}}\, \hat\phi_i\, (I_{[i=j]} + I_{[i \ne j]} \hat\phi_j)\, (\hat\beta_{m(i)} + \frac{\hat e_i}{\hat\phi_i}\, \hat c_{m(i)})\, (\hat\beta_{m(j)} + \frac{\hat e_j}{\hat\phi_j}\, \hat c_{m(j)}) \tag{14}$$

3. Large Sample Properties of Variance Estimators

We next consider the large-sample behavior of the survey estimator $\hat{Y}$ and its associated variance estimators under a broad but not fully general range of assumptions incorporating misspecification of the adjustment model.

(C.0) Sampling within each half-PSU $\mathcal{U}_{k,H}$ is simple random sampling of n_{kH} units out of N_{kH}. For simplicity, let $f \equiv n_{kH}/N_{kH}$ be the same for all k, H, and f be small, so that factors $1-f$ can be replaced by 1.

(C.1) There is a partition of $\mathcal{U}$ into cells $B_1, B_2, \ldots, B_L$ such that the true response probabilities $Er_i = \phi_i^{-1}$ are piecewise constant on each B_l, i.e.

$$\phi_i^{-1} = P(r_i = 1) \equiv \rho_l \qquad \text{whenever} \qquad i \in B_l, \;\; 1 \le i \le N, \;\; 1 \le l \le L$$

For each $i \in \mathcal{U}$, denote by $l = l(i)$ the unique index $l = 1, \ldots, L$ for which $i \in B_l$.

(C.2) As $n, N \to \infty$, the numbers L and M of cells and the number K of PSU's are fixed, and the sizes of the true cells B_l and working cells C_m and their intersections relative to the frame population have limits for all $1 \le l \le L$, $1 \le m \le M$, $1 \le k \le K$, $H = 1, 2$,

$$\lim_{n,N\to\infty} \; N^{-1} \sum_{i \in \mathcal{U}} I_{[i \in U_{k,H} \cap B_l \cap C_m]} \;\equiv\; \nu(l, m, k, H)$$

Our method of studying the large-sample limit limit of variance estimators is to express limiting variances in terms of the joint discrete mass function $\nu(l, m, k, H)$ of the random indices l, m, k, H. For example, under Assumptions (C.1)-(C.2), the limits $c_{m(i)} = 1 + \lambda_0' x_i$ of the ratio-adjustment factors $\hat{c}_{m(i)} \equiv 1 + \hat{\lambda}' x_i$ are

$$c_m \;=\; \lim_{N\to\infty} \frac{\sum_{i \in C_m \cap \mathcal{S}} 1/\pi_i}{\sum_{i \in C_m \cap \mathcal{S}} r_i/\pi_i} \;=\; \frac{\sum_{i \in C_m \cap U} 1}{\sum_{i \in C_m \cap U} \phi_i^{-1}} \;=\; \frac{\sum_{k,H,l} \nu(l, m, k, H)}{\sum_{k,H,l} \rho_l \, \nu(l, m, k, H)} \tag{15}$$

a formula with the nice conceptual interpretation $c_m = 1/E_\nu(\rho_l \,|\, m)$.

For simplicity in developing formulas for large-sample limits of variance estimators, we next restrict the way in which attributes can vary within PSU's and cells by assuming as $N \to \infty$:

(C.3) The attributes y_i for $i \in \mathcal{U}$ behave as independent random variables with uniformly bounded third absolute moments, and with variances σ^2 and means μ_k depending only on the PSU k to which i belongs.

(C.4) The replicates $f_{it} = 1 + 0.5\,(-1)^H a_{kt}$ for $i \in \mathcal{U}_{kH}$ satisfy for all working-cell indices m,

$$N^{-1} \sum_{i \in C_m} (1 - \phi_i^{-1})\,(f_{it} - 1) \;\to\; 0 \qquad \text{and} \qquad N^{-1} \sum_{i \in C_m} \phi_i^{-1}\,(f_{it} - 1) \;\to\; 0 \qquad \text{as} \qquad N \to \infty$$

Remark 1 *Assumption* **(C.4)** *holds quite generally but does depend on the precise mechanism used to split PSU's and attach Hadamard-matrix columns* $\{a_{kt}\}_{t=1}^{R}$ *to them. If splitting were done by attaching to each individual* i *within stratum* k *an independent binary label* $H = 1, 2$ *(with equal probabilities), so that the sequence of terms* $(-1)^H$ *in the definition of* f_{it} *form a sequence of iid random signs, then* **(C.4)** *holds by the Law of Large Numbers.*

Large-sample formulas for estimators of variance of $\hat{Y}$ simplify interestingly in two special cases. The first requires that the proportional decomposition of cells and strata is the same in the two half-strata indexed by H :

(Case A) For all k, l, m, $\quad \nu(l, m, k, 1) = \nu(l, m, k, 2)$.

Since approximate conditional independence under ν of the true-cell index l from the half-PSU indices (k, H) given the working adjustment-cell index m turns out to explain much of the remarkably good expected behavior of the BRR variance estimators under misspecified cell-based weighting adjustments for nonresponse, we define:

(Case B) For all k, l, m, H, $\quad \nu(l, m, k, H)/\sum_{l'} \nu(l', m, k, H)$ does not depend on (k, H).

3.1 Limiting Variance Formulas based on (9) and (13)

For any attribute z_i, denote respectively by $\bar{z}_{kH}$ and $s^2_{kH,z}$ the within-U_{kH} population mean and variance

$$\bar{z}_{kH} \;=\; (N_{kH})^{-1} \sum_{i \in U_{kH}} z_i \quad , \qquad s^2_{kH,z} \;=\; (N_{kH}-1)^{-1} \sum_{i \in U_{kH}} (z_i - \bar{z}_{kH})^2$$

Also recall that in **(C.0)**, the sizes N_{kH} are all large enough that $N_{kH}/(N_{kH}-1) \;\approx\; 1$.

Under assumptions **(C.0)-(C.3)**, the limiting forms of regression coefficients $\hat{\beta}$ and residuals $\hat{e}_i$ are

$$\beta^0_m \;=\; \text{P-}\lim_N \hat{\beta}_m \;=\; \sum_{k,H,l} \nu(l,m,k,H)\, \rho_l\, \mu_k \Big/ \sum_{k,H,l} \nu(l,m,k,H)\, \rho_l \tag{16}$$

$$e^0_i \;=\; y_i \;-\; \text{P-}\lim_N \hat{\beta}' x_i \;=\; y_i \;-\; \beta^0_{m(i)} \tag{17}$$

and it is shown in Slud and Thibaudeau (2008, Appendix) also that

$$(f/N)\, V(\hat{Y}) \;\approx\; \sigma^2 \sum_{l,m,k,H} c_m\, \nu(l,m,k,H) + \sum_{l,m,k,H} \nu(l,m,k,H) \left[(\mu_k - \beta^0_m)^2\, \rho_l (1-\rho_l)\, c^2_m + (\beta^0_m + (\mu_k - \beta^0_m)\rho_l\, c_m)^2 \right]$$

$$- \sum_{k,H} \nu(k,H) \Big\{ \sum_{l,m} (\beta^0_m + (\mu_k - \beta^0_m)\rho_l c_m)\, \nu(l,m|k,H) \Big\}^2 \tag{18}$$

and that the limit of the Särndal and Lündstrom (2005) variance (9), with $1-f$ replaced by 1, is

$$V_{SL}(\hat{Y}) \;\approx\; (N/f) \sum_{l,m,k,H} \Big[\sigma^2\, c_m \;+\; (c_m - 1)\, (\mu_k - \beta_m)^2 \Big]\, \nu(l,m,k,H) \tag{19}$$

3.2 Formulas for Expectation of BRR Variance Estimators

To study the average of $V_{\rm BRR}(\hat{Y})$, we start from the asymptotically equivalent form (5) displayed above. This approximation is very good in practice, as is documented via simulation by Slud and Thibaudeau (2008), with error at most one or two percent, when K is large and R is at least $K/4$.

Slud and Thibaudeau (2008, Prop. 4) find a rather complicated but explicit large-sample limit for the expression

$$(f/N)\, E(V_{\rm BRR}) \;-\; (fN)\, V_* \tag{20}$$

where

$$V_* \;\equiv\; \sum_k \Big(\sum_{l,m} (\beta^0_m + \rho_l c_m (\mu_k - \beta^0_m))(\nu(l,m,k,1) - \nu(l,m,k,2)) \Big)^2 \tag{21}$$

In special cases **(A)** and **(B)**, the complicated general limiting form for (20) of the $V_{\rm BRR}$ formula is simpler and interpretable. Numerical comparisons of this formula with (9) and (13) will be given in Section 4.

3.3 Simplifications under Cases (A)–(B)

Conditions **(C.0)-(C.4)**, which led to the expected $V_{\rm BRR}(\hat{Y})$ formula Slud and Thibaudeau (2008, Prop. 4), do allow the adjustment cells to be misspecified. In our notation, this is reflected by the products $\rho_l\, c_m$ being different from 1 when the random index-quadruple (l,m,k,H) has probability mass function ν. The other mechanism which affects differences $(f/N)\,(E(V_{\rm BRR}) - V(\hat{Y}))$ or $(f/N)\,(E(V_{\rm BRR}) - \hat{V}(\hat{Y}))$ is imbalances violating **Case (A)**. The main result of this paper is that a combination of both effects can result in meaningfully large discrepancies $E(V_{\rm BRR}) - V(\hat{Y}))$. The following theoretical results, proved in Slud and Thibaudeau (2008), go far toward explaining why both mechanisms must be operating in order for large discrepancies to arise.

Under **Case (A)**, it is easy to check that the quantity V_* in (21) is identically 0, and the limiting form of the expected BRR variance is given in the large-superpopulation limit by

$$\lim (f/N)\, E(V_{\rm BRR}) \;=\; \lim (f/N)\, V(\hat{Y}) \;=\; \quad (18)$$

but these limits are not generally exactly equal to (19).

Case (B) says that in the large-superpopulation limit, under the joint distribution $\nu(l,m,k,H)$, the index l is conditionally independent of (k,H) given m. It follows that $\beta_m^0 = \beta_m$, and in the limit,

$$\lim (f/N)\, V(\hat{Y}) \;=\; \lim (f/N)\, V_{SL}(\hat{Y}) \;=\; (19)$$

while

$$\lim \Big\{ (f/N)\, E(V_{\mathrm{BRR}}) - (fN)\, V_* \Big\} \;=\; \sum_{l,m,k,H} \Big(\sigma^2 c_m + (c_m - 1)(\mu_k - \beta_m^0)^2 \Big)\, \nu(l,m,k,H) + \sum_{l,m,k,H} (c_m - 1)(\mu_k - \beta_m^0)^2 \cdot$$

$$\cdot \Big\{ (\nu(k,1) - \nu(k,2))^2\, \nu(l,m,k,H) \;+\; (\nu(l,m,k,1) - \nu(l,m,k,2))\, (\nu(k,1) - \nu(k,2)) \Big\}$$

4. Numerical Comparisons

We consider now a brief numerical study in the setting of **(C.0)-(C.4)** comparing the large-sample theoretical formulas (19) for $V_{SL}(\hat{Y})$, (18) for the true variance $V(\hat{Y})$, and the limiting expected BRR variance formula of Prop. 4 of Slud and Thibaudeau (2008). Our objective is to understand the likely magnitudes of relative bias of $\hat{V}_{\mathrm{BRR}}$ for $V(\hat{Y})$ in terms of qualitative restrictions on the array $\nu(l,m,k,H)$ of limiting large-sample proportions of the population concentrated in $\mathcal{U}_{kH} \cap B_l \cap C_m$.

We illustrate several numerical examples with $\nu(l,m,k,H)$ arrays initially defined to satisfy **(A)** and nearly satisfying **(B)** and then violating **(A)** more and more strongly.

In the first set of examples, we fix the following parameters:

$$\sigma^2 = 0.2 \;\;,\quad L = M = 10 \;\;,\quad N = 10^6 \;\;,\quad f = 0.004 \;\;, \{\rho_l\}_{l=1}^{10} = 0.6 + (0, \ldots, 9) \cdot 0.4/9$$

so that the average response rate is always 0.8. For an integer $q = 4$ or 16,

$$K \;=\; 5q \quad , \qquad \{\mu_k\}_{k=1}^{5} \quad = \quad q \text{ concatenated copies of } \qquad (1.5,\ 1.75,\ 2,\ 2.25,\ 2.5)$$

so that $\bar{\mu}$ is close to 2, and the attributes y_i have coefficient of variation roughly 0.25. The expansion factors q, when applied to the arrays $\nu(l,m,k,H)$ defined in the next paragraph, determine that the number of indices k is inflated by the factor q, with each of the sub-arrays $\{\ \nu(l,m,k,H)\}_{l,m,H}$ copied q times and pasted together to form a new $L \times M \times (5q) \times 2$ array.

Each of the ν arrays we used was generated along to the following scheme. The 'true' cells B_l were defined to have equal size, here $\nu(l,\cdot,\cdot,\cdot) = \sum_{k,m,H} \nu(l,m,k,H) = 1/L = 0.1$ for each $l = 1, \ldots, 10$. Then the bivariate *confusion matrix* $\nu(l,m,\cdot,\cdot) = \sum_{k,H} \nu(l,m,k,H)$ was specified so that the conditional entries $\nu(m|l) \equiv \nu(l,m,\cdot,\cdot)/\nu(l,\cdot,\cdot,\cdot)$ were proportional to $\exp(-c\, l^a \cdot |l - m|)$, where c is a constant and a was 1, 0.5, or 0. Next, the ν conditional distribution of k given (l,m) was specified to depend only on k, l proportionally to $\exp(\beta\, k\, l)$ where the constant β was taken as either .01 or .03. Note that in this formulation, k and m are taken conditionally independent given m. To complete the specification, $\nu(H|l,m,k) \equiv \nu(l,m,k,H)/(\nu(l,m,k,1) + \nu(l,m,k,2))$ is defined as 0.5 wherever we wanted to maintain perfect balance as in **Case (B)**, and otherwise as a single array (generated once only for each example ν) of independent identically distributed variates $u_{k,l,m} \sim \mathrm{Uniform}(\frac{1}{2}(1 - \omega), \frac{1}{2}(1 + \omega))$, where $\omega \in (0, 0.5)$ is an input parameter, used below to quantify imbalance.

Four different arrays generated in this way can be described in terms of two summary statistics:

`SDcond` = average over (k,H) of $\mathrm{SD}(\nu(l|m,k,H))$ (*which measures violation of* **Case (B)**)

`missp` = Misspecification $(\sum_{l,m} \nu_*(l,m)\, (\rho_l\, c_m - 1)^2)^{1/2}$ (*measuring misspecification of cells*)

The value `SDcond` was defined for each array as the average over (l,m) of the standard deviations for fixed (l,m) of the vectorized array of numbers $\{\nu(H|l,m,k)\}_{k,H}$, and would be 0 under **Case (B)**. The quantity `missp` depends only on the joint distribution of (l,m) under ν, and therefore is unaffected by ω and the size factor q used to vary the number of PSU's $K = 5q$. For the four example ν arrays used in constructing Table 1 below, the resulting values of `missp` are as follows:

Example ν array	1	2	3	4
`missp` value	0.159	0.116	0.121	0.069

Table 1 displays a series of comparative values calculated and simulated for the large-sample variances of the survey estimator $\hat{Y}$. Columns 1–4 of the Table specify an example, i.e., the specific $\nu(l, m, k, H)$ array used before multiplicative perturbation by Unif$[1-\omega, 1+\omega]$ variates, as described above. The column entries q denote in each row the multiple of 5 used to define the number K of PSU's in the example. The `SDcond` column is the computed descriptive statistic `SDcond` for the Example in each row. Columns 5–7 of the Table are the variances multiplied by f/N , respectively $E(V_{\rm BRR}(\hat{Y})$ as given in Prop. 4 of Slud and Thibaudeau (2008) and V_{SL} given in (11) and $V(\hat{Y})$ given in (18). Unlike columns 5–7, which are the result of theoretical calculation, the final columns of the Table are compiled from the results of a simulation study conducted under the survey design. The stratified survey of sample-size $n = 4000$ specified in each row of the Table was simulated independendently 1000 times. Then columns 8–10 of Table 1 contain respectively the empirical average over the simulation-iterations of formula (18), the empirical standard deviation of the estimator $\hat{V}_{SL}(\hat{Y})$ given in formula (10), and the empirical standard deviation of the BRR variance estimator (4).

For the most part, the BRR formula variances (column 5) in Table 1 are seen to be close to the true ones (column 7), especially in examples with $\omega = 0$, because we have formulated as a base case for each ν array example a setting where **Case (A)** holds precisely and **(B)** holds approximately. (The quantities `SDcond` have intentionally been allowed to remain rather small.) However, as ω increases and exact half-PSU balance is violated, it is not hard to generate examples showing extremely large bias for the BRR variance estimators. In fact, a well-designed survey should not show anything like as much half-PSU imbalance as is embodied (through $\omega = 0.1$) in randomly selecting a probability in the range 0.45–0.55 as a PSU-splitting fraction. When $\omega < 0.05$, it is quite hard to generate examples with numbers of strata and cells like those here in which the bias between BRR and $V(\hat{Y})$ is more than a few percent.

The main tendency of these numerical comparisons has been to point out

- that the expected BRR variance is extremely close to the V_{SL} and $V(\hat{Y})$ variances in almost all practical cases under **Case (B)**, and
- when there is even the slightest violation of **Case (B)**, then the differences between $EV_{\rm BRR}$ and $V(\hat{Y})$ can become interestingly large (in a way which depends on the adjustment cell misspecifications and PSU differences in cell proportions), with the former generally larger.

The simulation gives us three specific conclusions. First, the close agreement between the columns 7 and 8 of Table 1 corroborates the theortical approximation (18) to $V(\hat{Y})$, and the empirical average of (4) agreed similarly with the theoretical formula. Second, the approximation of (4) by (5) was seen to be extremely close (within 1% almost without exception). Finally, we can see from the final two columns of the Table that the variance of the BRR variance estimator can be quite high compared to other estimators of good quality, like $\hat{V}_{SL}$.

5. Summary & Discussion

This paper has studied the bias of Replication-based variance estimators only in the simplest possible setting: that of a survey in which PSU's (whether sampled or self-representing) are split in order to generate a system of replicate weight factors (as in Fay 1984) $1 \pm (0.5)$ which are constant on each half-PSU. In that setting, we have introduced a pseudo-randomization model for nonresponse and have elaborated a set of superpopulation model assumptions under which

(1) survey attributes behave like independent variates, identical except that means may differ across PSU's;

(2) nonresponse probabilities are constant within a fixed finite set of response cells, and the proportion of the frame population within true cell l, working adjustment cell m, PSU k and half-PSU index H has a superpopulation limit; and

(3) SRS sampling is done within PSU's.

Of these Assumptions, (1) is not realistic but rather represents a useful extreme where there are no relationships between attributes, nonresponse, and the demographic covariates defining adjustment cells, to lead to bias in the survey-weighted attribute totals. The restrictions (2) and (3) are made purely for reasons of theoretical tractability, and could easily be generalized. Our results show in this setting:

(a) Even when the adjustment cells are misspecified, the BRR variance estimators are generally unbiased when the PSU by response and adjustment cells are split evenly, in a probabilistic or superpopulation sense.

Table 1: Variances multiplied by f/N for $\hat{Y}$ in examples of survey designs. Columns 5–7 are theoretical calculated values given by formulas described in the text, while columns 8–10 report related simulation results.

				Calculated			**Simulated**		
ν Array	q	ω	SDcond	EV_{BRR}	V_{SL}	$V(\hat{Y})$	$\overline{VY}$	VB.sd	VB.sd
1	4	0	.0028	1125	1125	1130	1119	346	55
	16	0	.0008	1125	1125	1130	1124	192	61
	4	.1	.0032	1256	1125	1130	1116	390	56
	16	.1	.0008	1158	1125	1130	1128	183	60
2	4	0	.0026	1146	1146	1160	1144	374	56
	16	0	.0007	1147	1146	1159	1152	182	62
	4	.1	.0030	1374	1146	1160	1151	414	57
	16	.1	.0008	1204	1146	1160	1146	203	60
3	4	0	.0063	1141	1140	1183	1144	373	59
	16	0	.0017	1141	1140	1183	1146	187	61
	4	.1	.0064	1445	1140	1183	1144	484	60
	16	.1	.0017	1217	1140	1183	1139	209	63
4	4	0	.0063	1157	1157	1174	1148	381	60
	16	0	.0017	1157	1157	1174	1165	187	62
	4	.1	.0065	1844	1157	1174	1155	608	59
	16	.1	.0017	1329	1157	1174	1154	238	63

(b) When imbalances appear in the way the PSU's and true and working response cells are split, biases in the BRR variance estimators can be large, usually in the direction of inflating the variance.

(c) To an extent which must be clarified in further research, the BRR variance estimators do **not** converge in the large-sample superpopulation limit to the constant values of the true variances, but rather have asymptotic distributions showing residual variability that can be large when the adjustment cells are misspecified.

There are a few lessons to draw from the paper. When the mechanism of splitting into half-PSU's cuts evenly across response cells, as is likely to be true by the alternate-indexing of systematic samples used in the Survey of Income and Program Participation (SIPP), then BRR provides reliably unbiased estimators, and this research supports the use of such estimators. Second, there is some need for empirical research on whether split-PSU imbalances that might cause BRR variance estimation biases really do arise in practice. Finally, it should not be at all difficult to devise random mechanisms of splitting PSU's which lead to negligible imbalances, in which case the BRR estimators remain very attractive for their simplicity of interpretation in public-use files, the ease of programming to make use of them, etc. A fair coin-toss allocation to H within PSU is one way to accomplish this, but a better way would be to do some sort of balanced block randomization within covariate-defined cells. The allocation of sampled units to H=1,2 could also be done after sampling, taking some demographics into account for the responders.

REFERENCES

Brick, M., Morganstein, D. and Valliant, R. (2000), "Analysis of Complex Sample Data using Replication," *WESTAT Technical Paper*, July 29, 2000.

Fay, R. (1984), "Some properties of estimators of variance based on replication methods," in *ASA Proceedings*, Survey Res. Meth. Section, pp. 495-500. Alexandria, VA: American Statistical Association.

Fay, R. (1989)," Theory and application of replicate weighting for variance calculations," *ASA Proceedings*, Survey Res. Meth. Section, pp. 212-217. Alexandria, VA: American Statistical Association.

Kish, L. and Frankel, M. (1970), "Balanced repeated replications for standard errors. *Jour. Amer. Statist. Assoc.*, 65, 1071-1094.

Kim, Jae Kwang and Kim, Jay J. (2007), "Nonresponse weighting adjustment using estimated response probability," *Canad. Jour. Statist.*, 35, 501-514.

Oh, H. and Scheuren, F. (1983), "Weighting adjustment for unit nonresponse," in: *Incomplete Data in Sample Surveys*, vol. 2, Eds. Madow, W., Olkin, I. and Rubin, D., 143-184. New York: Academic Press.

Särndal, C.-E. and Lündstrom, S. (2005), *Estimation in Surveys with Nonresponse*. Wiley: Chichester.

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

Slud, E. and Thibaudeau, Y. (2008), "BRR versus inclusion-probability formulas for variances of nonresponse-adjusted survey estimates", *Techincal Report*, Census Bureau Staistical Res. Div.

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