REPLICATE VARIANCE ESTIMATION FOR THE NATIONAL SURVEY OF PARENTS AND YOUTHS

Louis Rizzo and David Judkins
Louis Rizzo, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: multi-stage sampling, jackknife, finite population correction, balanced repeated replication, Durbin-Brewer, Yates-Grundy

1. Introduction

Replicate variance estimation has become a standard technique for generating consistent variance estimators for a wide range of estimators from complex sample surveys. One of its greatest virtues is to provide for an easy computation of variances of nonlinear functions of weighted means and totals[1], such as ratios, regression coefficients, etc., without special formulas for each application. A longstanding problem for replicate variance estimation is the incorporation of finite population corrections and nontrivial joint probabilities of selection. In particular in multi-stage surveys, capturing nontrivial joint probabilities in the first stage of selection as well as nontrivial second-stage finite population corrections has not been easy.

Kalton (1979) for example studies the circumstances under which one can approximate an epsem multi-stage design by a simple random sample or a stratified random sample of 'ultimate clusters', which are final aggregations within the first-stage sample units of sampled second-stage sample units. If this approximation is valid, then the sample variance of the multi-stage estimator can be approximated by a sum of squares of ultimate cluster means or totals. The second-stage finite population correction gets incorporated appropriately without including it explicitly (it is appropriately reflected in the sums of squares of the totals). It is also easy then to generate replicate weights to reproduce these sums of squares, generating replicates at the first stage of selection only (i.e., perturbing weights at the level of first-stage sample units only). However, generally this approximation only works if the first-stage finite population corrections are small.

Rao and Wu (1988) solve the problem of having nontrivial finite population corrections at multiple stages of selection by providing consistent bootstrap estimators for multi-stage designs. Appropriate scaling factors are introduced for each stage of a 'two-stage bootstrap', replicating in this way the original sampling process with its divergent finite population corrections. For jackknife or BRR variance estimation however, we prefer to have only one set of replicate weights that cover all stages of selection in an appropriate way, for the sake of operational simplicity. Flyer (1987) presents BRR replicate estimates which succeed in capturing nontrivial first-stage finite population corrections, but uses an extra randomization in defining replicate weights, which adds a component of variability to the variance of the variance estimator. Fuller (1998) presents a jackknife estimation procedure for regression estimators in two-phase samples, a related problem.

In this paper, replicate weights are developed which perturb the first-stage sample units only, based on a jackknifing procedure and 'eigenmatching': a procedure developed by Fay, and used recently by Fuller (1998). This is designed to match a sum of squares at the first-stage sample unit level. A second set of replicate weights, which perturb at the level of second-stage sample units, is based on balanced repeated replication (BRR), using Fay's method for BRR (see for example Judkins 1990). The combined set of replicate weights produces a sum of squares which matches the unbiased variance estimator of the total including all stages, appropriately including joint probabilities at the first stage of selection and finite population corrections at the second stage of selection. As far as we know, no-one has done this and published the results.

Section 2 gives an overview of the relevant aspects of the sample design for which the method was developed. Section 3 provides the unbiased variance estimator for totals for this sample design: the 'matching template' for the replicate variance estimator. Section 4 is a development of the eigenanalysis for the eigenmatching of the PSU-level sums of squares component of the unbiased variance estimator. Section 5 described the replicate variance estimator based on the eigenanalysis. Section 6 documents the replicate variance estimator developed to match the complement sums of squares of the unbiased variance estimator. Finally, Section 7 provides a discussion of what we accomplished.

2. Overview of the NSPY Sample Design

The National Survey of Parents and Youth (NSPY) was a national sample of youths (ages 9 to 17 at baseline) and their parents designed to evaluate the National Youth Anti-Drug Media Campaign (NYAMC). The surveys centered on questions about youth and parent exposure to the media campaign, youth attitudes about drug use, parental attitudes about drug use by their children, and parental actions to prevent such use.

A total of 90 PSUs were sampled for NSPY. These 90 PSUs were selected from a larger 'master sample' of 100 PSUs that Westat drew in the early 1990s for multiple use by a range of possible surveys. This original master sample included 24 certainty PSUs (i.e., PSUs that were included in the sample with probability 1), and 76 noncertainty PSUs (i.e., PSUs selected with probability strictly less than 1) selected from 38 strata (two sampled PSUs per stratum). PSU selection was independent across these 38 strata. The

[1] Weighted by the sample weights, which include the inverse of the probability of selections for the levels of the sample design, nonresponse adjustments, and possibly calibration adjustments.

Durbin-Brewer sampling method was utilized to select each stratum sample (Durbin 1967, Brewer 1963).

Under the Durbin-Brewer methodology, a measure of size z_{ti} is assigned to each PSU in each stratum[2]. The measures of size need to be proportionately resized to sum to 1 for each stratum. For the PSU master sample, these z_{ti} values were proportional to the 1990 PSU population. Inclusion probabilities p_{ti} are set equal to $2*z_{ti}$ for each PSU. The actual details of how the sample is drawn within each stratum under this sample design is discussed for example in Section 9A.8 of Cochran 1977: for variance estimation purposes here we need only the joint probabilities of inclusion $p_{ti,tj}$ induced under this design for pairs of PSUs *ti* and *tj* within one stratum. These joint probabilities are as follows (see for example Equations 9A.45 and 9A.47 in Cochran 1977):

$$p_{ti,tj} = \frac{z_{ti}z_{tj}}{D_t}\left(\frac{1}{1-2z_{ti}}+\frac{1}{1-2z_{tj}}\right) \text{with } D_t = \sum_{i=1}^{N_{ts}} \frac{z_{ti}\left(1-z_{ti}\right)}{1-2z_{ti}},$$

where N_{ta} is the number of PSUs on the frame for first-stage stratum a in superstratum t. It is important to note that with the joint probabilities defined in this way that $\Delta_{ti,tj} = p_{ti}p_{tj} - p_{ti,tj}$ is always greater than or equal to 0 (see the discussion in Cochran 1977, p. 262, following Equation 9A.47).

The 90 NSPY PSUs were selected from the 100 master sample PSUs in the following way. The twenty largest (in terms of 1990 population) master sample certainty PSUs were selected with certainty into the NSPY PSU sample. The remaining 4 master sample certainty PSUs with the 76 master sample noncertainty PSUs were then placed into 10 'superstrata', with eight PSUs per superstratum. The 10 superstrata are subscripted as $t=1,...,10$. A sample of 7 PSUs was drawn with equal probability and without replacement from the 8 master sample PSUs within each superstratum: these 7 PSUs comprise the NSPY PSU sample for the superstratum. So within each superstratum, there were three original strata with two sampled PSUs each, and one original stratum with just a single sample PSU. PSUs will be subscripted by *ti* (the *i*-th PSU in the *t*-th superstratum). The 20 certainty PSUs are placed in superstratum t=11.

For the 70 noncertainty PSUs, the probability of inclusion π_{ti} for each of these PSUs into NSPY is

$$\pi_{ti} = \frac{7}{8}p_{ti}$$

where p_{ti} is the probability of selection of the PSU into the master sample[3]. The joint probabilities of inclusion are written as $\pi_{ti,tj}$. For PSUs i and j in different superstrata t and u, sampling is independent, so that $\pi_{ti,uj}$ equals $\pi_{ti}\pi_{uj}$. For PSUs i and j in the same superstratum t, the unconditional joint probability of selection is as follows:

$$\pi_{ti,tj} = \begin{cases} \frac{6}{8}p_{ti}p_{tj} & ti,tj \text{ in different master sample strata} \\ \frac{6}{8}p_{ti,tj} & ti,tj \text{ in same master sample stratum} \end{cases}$$

The second stage of selection (conditional within PSUs for the noncertainty PSUs, and unconditional for the certainty PSUs) is of segments within PSUs, followed by selection of dwelling units within segments, and youth within households. The sampling process of segments was a probability proportionate to a measure of size systematic sampling of segments created from the 1990 Census block frame for each PSU[4]. The segments were designed to include 60 households, though in practice there was variation around the target figure.

For the purposes of explicating the NSPY variance estimation approach, we approximate the within-PSU sampling process as follows. For each PSU *ti* there is a total of S_{ti} second-stage strata, with a sample size of m_{tis} clusters drawn from each stratum (m_{tis} is generally 2, but is sometimes 3 for at most one stratum in each PSU). These second-stage strata are defined based on the frame order for segment selection within the PSU. The second-stage strata will be subscripted as *tis* ($s=1,...,S_{ti}$), with M_{tis} defined as the number of clusters in the population in the stratum (clusters within second-stage strata are subscripted as *tisc*, $c=1,...,M_{tis}$ for the population and $c=1,...,m_{tis}$ for the sample). We will write f_{tis} as the sampling fraction for the stratum: $f_{tis} = m_{tis}/M_{tis}$. Note that f_{tis} is small when π_{ti} is large and vice-versa since the segment sample was selected in such a manner as to make the overall sample close to a self-weighting sample.

3. Variance Estimators for Totals Under the Sample Design

Suppose we have a variable Y which is a cluster-level count of persons with a particular characteristic. The overall population total of Y is as follows:

$$Y = \sum_{t=1}^{11} Y_t = \sum_{t=1}^{11}\sum_{i=1}^{N_t} Y_{ti} = \sum_{t=1}^{11}\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\sum_{c=1}^{M_{tis}} Y_{tisc}$$

The Y_t ($t=1,...,10$) are the population totals for the superstrata, with Y_{11} ($t=11$) corresponding to the population

[2] The subscripts *ti* indicate PSU *i* in 'superstratum' *t*. The superstrata are defined below: they consist of four PSU strata each.

[3] Note that this PSU subsampling procedure induced more uniform weights then the alternative procedure of purposely selecting an original stratum to lose a sample PSU, but that the procedure does induce a between-stratum variance component which must be reflected in the variance estimates.

[4] There was also a selection from segments reflecting new construction since 1990.

total for the certainty PSUs. N_t is the total number of PSU's in the population in superstratum t, $t=1,\ldots,10$, with N_{11} equal to the number of certainty PSUs: 20. Y_{ti} is the population total for the PSU ti (the i-th PSU in the t-th superstratum), M_{ti} the total number of clusters in each frame PSU ti. M_{tis} is the total number of clusters in second-stage stratum s, with S_{ti} strata for each PSU ti. Y_{tisc} is the cluster total for each cluster $tisc$. For example, if Y is the number of 9 to 11 year olds who have smoked marijuana, then Y_{tisc} is the total number of 9 to 11 year olds in the interviewed households from the sampled segment who have smoked marijuana.

The estimator of Y from the final sample can be written either as a weighted summation of sample PSU totals or as a weighted summation of sample cluster totals (with π_{ti} equal to 1 for the certainty PSUs ($t=11$)):

$$\hat{Y} = \sum_{t=1}^{11}\sum_{i=1}^{n_t}\frac{\hat{Y}_{ti}}{\pi_{ti}} = \sum_{t=1}^{11}\sum_{i=1}^{n_t} w_{ti}\hat{Y}_{ti} = \sum_{t=1}^{11}\sum_{i=1}^{n_t}\sum_{s=1}^{S_{ti}}\sum_{c=1}^{m_{tis}} w_{ti}w_{tis}\hat{Y}_{tisc}$$

$$\text{with } w_{ti} = \frac{1}{\pi_{ti}} \quad \text{and} \quad w_{tis} = \frac{1}{f_{ti}}$$

$\hat{Y}_{tisc}$ is the unbiased estimator of Y_{tisc} based on the youth and parent sample within the sampled households. The quantity f_{ti} is equal to the second-stage stratum tis sampling fraction m_{tis}/M_{tis}, which under the sample design is constant within each PSU[5].

The variance of $\hat{Y}$ can be computed as follows (see for example Chapter 11 in Cochran 1977):

$$Var\left(\hat{Y}\right) = \sum_{t=1}^{10}\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj} - \pi_{ti,tj}\right)\left(\frac{Y_{ti}}{\pi_{ti}} - \frac{Y_{tj}}{\pi_{tj}}\right)^2 + \sum_{t=1}^{11}\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}f_{ti}^2}S_{tisy}^2,$$

where S_{tisy}^2 is the population variance of the cluster totals Y_{tisc}. Generalizing Equation 11.44 in Cochran 1977, the following variance estimator is shown in the appendix to be an unbiased estimator of this variance:

$$v\left(\hat{Y}\right) = \sum_{t=1}^{10}\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\pi_{ti}\pi_{tj} - \pi_{ti,tj}}{\pi_{ti,tj}}\left(\frac{\hat{Y}_{ti}}{\pi_{ti}} - \frac{\hat{Y}_{tj}}{\pi_{tj}}\right)^2 + \sum_{t=1}^{11}\sum_{i=1}^{n_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}(1-f_{ti})}{\pi_{ti}f_{ti}^2}s_{tisy}^2$$

$$\text{with } s_{tisy}^2 = \frac{1}{m_{tis}-1}\sum_{c=1}^{m_{tis}}\left(y_{tisc} - \bar{y}_{tis}\right)^2$$

$$\text{and } \bar{y}_{tis} = \frac{1}{m_{tis}}\sum_{c=1}^{m_{tis}} y_{tisc}$$

This variance estimator can be decomposed into two terms which provide the 'matching templates' for the replicate variance estimators:

$$v\left(\hat{Y}\right) = v_1\left(\hat{Y}\right) + v_2\left(\hat{Y}\right) \quad \text{with}$$

$$v_1\left(\hat{Y}\right) = \sum_{t=1}^{10}\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\pi_{ti}\pi_{tj} - \pi_{ti,tj}}{\pi_{ti,tj}}\left(\frac{\hat{Y}_{ti}}{\pi_{ti}} - \frac{\hat{Y}_{tj}}{\pi_{tj}}\right)^2$$

$$v_2\left(\hat{Y}\right) = \sum_{t=1}^{11}\sum_{i=1}^{n_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}(1-f_{ti})}{\pi_{ti}f_{ti}^2}s_{tisy}^2$$

Note that for the certainty PSU stratum ($t=11$) there is no first term, and for the second term π_{ti} is equal to 1 for $t=11$. It is important to note the $v_1\left(\hat{Y}\right)$ is *not* an unbiased estimator of the first term of $Var\left(\hat{Y}\right)$: its expectation exceeds the first term of Equation 6. However, $v_1\left(\hat{Y}\right)$ and $v_2(\hat{Y})$ together have as their expectation the variance $Var\left(\hat{Y}\right)$.

4. Eigenanalysis of the Variance Estimator

For NSPY variance estimation we developed a set of 100 replicate weights ($r=1,\ldots,100$). From these replicate weights 100 replicate estimates of $\boldsymbol{Y}$ can be computed:

$$\hat{Y}\left(r\right) = \sum_{t=1}^{11}\sum_{i=1}^{n_t}\sum_{s=1}^{S_{ti}}\sum_{c=1}^{2} w_{ti}(r)w_{tis}\left(r\right)y_{tisc}$$

The replicate variance estimator of the sampling variance of $\mathbf{Y}$ is then computed as follows:

$$v_{rep}\left(\hat{Y}\right) = v_{r_1}\left(\hat{Y}\right) + v_{r_2}\left(\hat{Y}\right) = b_1\sum_{r=1}^{60}\left(\hat{Y}(r) - \hat{Y}\right)^2 + b_2\sum_{r=61}^{100}\left(\hat{Y}(r) - \hat{Y}\right)^2$$

The factor b_1 is a constant greater than 1 whose selection will be discussed in the following sections. The replicate scheme was designed to produce a $v_{r1}\left(\hat{Y}\right)$ that

[5] Note that f_{ti} represents the sampling rate for both segment sampling and household sampling within segments: the second and third stages of selection.

matches $v_1\left(\hat{Y}\right)$ exactly when the $\hat{Y}$ is a weighted total, with $v_{r2}\left(\hat{Y}\right)$ matching $v_2\left(\hat{Y}\right)$ in expectation. Fay's method (Fay 1989) is used in the development of $v_{r1}\left(\hat{Y}\right)$, and is the subject of the development below. $v_{r2}\left(\hat{Y}\right)$ is based on a partially balanced repeated replication (BRR) approach: the constant b_2 is equal to 1/40.

The direct estimator $v_1\left(\hat{Y}\right)$ which we are attempting to match with a replicate estimator is equal to the summation of ten terms which have the form $\mathbf{x_{tw}}' \mathbf{C_t(s)}\, \mathbf{x_{tw}}$, where $\mathbf{x_{tw}}$ is the 7-vector of weighted sample PSU totals $\hat{Y}_{ti}/\pi_{ti}$ within superstratum t and the $\mathbf{C_t(s)}$ are a set of 7 by 7 matrices which are functions of the PSU joint probabilities of selection. Note that the criterion to apply Fay's method – that the matrix $\mathbf{C_t(s)}$ not be a function of y-values but only of probabilities of selection, sample sizes, etc. – is satisfied in this application.

The matrices $\mathbf{C_t(s)}$ are determined in fact from the method for selecting the PSUs within each superstratum, as described above. The on-diagonal elements of $\mathbf{C_t(s)}$ each correspond to one of the seven sampled PSUs in superstratum t; the off-diagonal elements each correspond to a pair of PSUs ti—tj. The first and second, third and fourth, and fifth and sixth values of i correspond to pairs of PSUs within the original PSU strata. The seventh value of i corresponds to the singleton PSU from the original PSU stratum sampled to have a PSU deselected from the PSU master sample (in the 7 in 8 subsampling process). Note that the off-diagonal elements 1-2, 3-4, 5-6 correspond to pairs of PSUs from the original master sample PSU strata, drawn together in the master sample Durbin-Brewer sampling process, and thereby represent complex joint selection probabilities. The remaining off-diagonal elements correspond to pairs of PSUs in differing original PSU strata, whose joint selection probabilities are entirely based on the 7 in 8 deselection process. It can be shown (available on request from the authors) that the eigendecomposition of each $\mathbf{C_t(s)}$ is as follows:

$$\mathbf{e}_{t1} = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}},\ 0,\ 0,\ 0,\ 0,\ 0\right]'$$
with eigenvalue $q_{t1} = (7/48) + \left(2 * a_{t1,t2}\right)$

$$\mathbf{e}_{t2} = \left[0,\ 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}},\ 0,\ 0,\ 0\right]'$$
with eigenvalue $q_{t2} = (7/48) + \left(2 * a_{t3,t4}\right)$

$$\mathbf{e}_{t3} = \left[0,\ 0,\ 0,\ 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}},\ 0\right]'$$
with eigenvalue $q_{t3} = (7/48) + \left(2 * a_{t5,t6}\right)$

$$\mathbf{e}_{t4} = \frac{1}{\sqrt{21}}\left[\frac{3}{2},\ \frac{3}{2},\ \frac{3}{2},\ \frac{3}{2}, -2, -2, -2\right]'$$
with eigenvalue $q_{t4} = 7/48$

$$\mathbf{e}_{t5} = \frac{1}{2}\left[1,\ 1, -1, -1,\ 0,\ 0,\ 0\right]'$$
with eigenvalue $q_{t5} = 7/48$

$$\mathbf{e}_{t6} = \frac{1}{\sqrt{6}}\left[0,\ 0,\ 0,\ 0,\ 1,\ 1, -2\right]'$$
with eigenvalue $q_{t6} = 7/48$

$$\mathbf{e}_{t7} = \frac{1}{\sqrt{7}}\left[1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1\right]'$$
with eigenvalue $q_{t7} = 0$

with

$$a_{ti,tj} = \left(\frac{\pi_{ti}\pi_{tj} - \pi_{ti,tj}}{\pi_{ti,tj}}\right) - \frac{1}{48}$$

Note that the last eigenvector (equal to a scaled multiple of the vector of 1's) has eigenvalue 0: the $\mathbf{C_t(s)}$ are not full-rank, but rank 6 matrices. The first three eigenvectors essentially correspond to 'within PSU stratum' differences: taking the squared differences of the PSU totals from within the original strata. The corresponding eigenvalues $(7/48) + \left(2 * a_{t1,t2}\right)$, $(7/48) + \left(2 * a_{t3,t4}\right)$, $(7/48) + (2 * a_{t5,t6})$ are based on the actual joint probabilities generated from the Durbin-Brewer sampling process. The values of these eigenvalues and details regarding their derivation for the NSPY study are available from the authors on request. Eigenvectors 4 through 6 (with eigenvalue 7/48) represent 'between PSU stratum' differences: taking the squared differences of PSU totals from differing original strata. The joint probability structure is entirely from the 7 in 8 subsampling process, resulting in the 7/48 eigenvalues[6]. Note that the three eigenvectors are not unique: any set of vectors spanning the three-dimensional linear space corresponding to the eigenvalue of 7/48 will also be valid eigenvectors.

5. Replicate Version of the Variance Component $v_1\left(\hat{Y}\right)$

Under the Fay approach, we found replicate weights that matched each of the ten components of $v_1\left(\hat{Y}\right)$, using the eigenanalysis described in the previous section. A total of six replicate weights which are subscripted as r=1,…,6 (a

[6] Note that the 1/48 comes from the computation of $\left(\pi_{ti}\pi_{tj} - \pi_{ti,tj}\right)/\pi_{ti,tj}$ in this case, which is ((7/8)*(7/8)-(6/8))/(6/8)

grand total of 60 replicates: six for each of the ten superstrata) will reproduce each $v_1\left(\hat{Y}_t\right)$.

The base weights w_{ti} for each PSU are given in Section 3. We define the vectors $\mathbf{w_t}$ for each superstratum t ($t=1,\ldots,10$), with each of these vectors having length 7, and having elements w_{ti}, ordered by PSU within superstratum ($i=1,\ldots,7$):

$$\mathbf{w_t} = \begin{bmatrix} w_{t1} & w_{t2} & * & * & w_{t7} \end{bmatrix}'$$

Similarly we can define a vector $\mathbf{y_t}$ of totals estimates across the sampled PSUs in superstratum t:

$$\mathbf{y_t} = \begin{bmatrix} \hat{Y}_{t1} & \hat{Y}_{t2} & * & * & \hat{Y}_{t7} \end{bmatrix}'$$

The estimators of the superstratum totals can be rewritten for the development below as follows:

$$\hat{Y}_t = \sum_{i=1}^{7} w_{ti}\hat{Y}_{ti} = \mathbf{w_t'y_t}$$

The following six replicate weight 7-vectors will achieve our goal of matching the quadratic form $v_1\left(\hat{Y}_t\right)$:

$$\mathbf{w_{tr}} = \mathbf{\Omega_t}\left(\mathbf{1_7} + \sqrt{\frac{q_{tr}}{b}}\mathbf{e_{tr}}\right) \qquad r = 1,\ldots,6$$

where the $\mathbf{e}_{tr}$, $t=1,\ldots,6$ are the six non-null eigenvectors of $v_1\left(\hat{Y}_t\right)$ (given in Section 4) the q_{tr} are the eigenvalues, which are the eigenvalues given in Section 3, $\mathbf{\Omega_t}$ is a 7 by 7 diagonal matrix with the w_{ti}, i=1,…,7 along the diagonal, and the factor b is a scaling value selected to ensure that all of the replicate weights are strictly positive. In order to do this, b must be greater than or equal to every eigenvalue q_{tr}. The maximum q_{tr} was 2.568, and b was set to this value.

6. Replicate Version of the Variance Component $v_2\left(\hat{Y}\right)$

We developed 40 partially balanced repeated replication (BRR) weights for r=61 through 100 using 'Fay's method' for BRR (see Judkins 1990), which gave us a consistent estimator of $v_2\left(\hat{Y}\right)$. Thus the forty sums of squares do not match $v_2\left(\hat{Y}\right)$ exactly, but approximately (there are nonzero but small cross-terms). This methodology is standard, and is not explicated in this paper. The detailed development of these replicate weights is available on request from the authors.

7. Discussion

The replicate weights represented in the replicate variance estimator $v_{rep}(\hat{Y})$ were designed to match the variance estimator $v(\hat{Y})$ for totals[7]. The replicate weights were carefully tailored to achieve this result: for $v_1(\hat{Y})$ the variance estimator was eigenanalyzed, with the replicate weights then constructed to replicate the eigenvectors and eigenvalues underlying the variance estimator, and for $v_2(\hat{Y})$ the matching process was a more standard scaled partially balanced repeated replication approach. As far as we know, this has not been done (and published) for a multi-stage survey. The most common approach in replication is to neglect the first-stage FPC while capturing the generally less important second-stage FPC. This is done by making the replicate weights reproduce unadjusted sums of squares at the first-stage level. These sums of squares generally capture the second-stage FPC, as the squares will be appropriately smaller where the second-stage FPC's are large. Another alternative is to incorporate first-stage FPC's into the replicate variance estimator, but this then distorts the second-stage FPC's. The NSPY approach we developed allows one to incorporate completely FPCs at both the first- and second-stage level. One caveat of this is as follows. The 'second-stage FPC' we matched was actually the complement of the product of sampling rates at two stages of NSPY selection (segment selection and household within segment selection), so that we did not capture completely the finite population corrections at every level of sampling, but only an aggregated finite population correction over two stages of sampling. The bias from this is essentially the difference between the target variance estimator $v\left(\hat{Y}\right)$ as given in Section 3, and the 'true' variance estimator accounting for all stages of selection exactly, the former being an approximation of the latter.

8. References

Brewer, K. W. R. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics 5*, 5-13.

Cochran, W. G. (1977). *Sampling Techniques, 3rd ed*. New York: John Wiley & Sons.

[7] The replicate variance estimator will be consistent (though not unbiased) for any smooth nonlinear function of sample means and totals as long as the perturbations are sufficiently small, and the function is sufficiently smooth. For example, to establish the consistency of the delete-one jackknife for $g(\bar{\mathbf{x}})$ for a k-vector of sample means $\bar{\mathbf{x}}$, a sufficient condition under simple random sampling is that the gradient of g needs to be nonzero and continuous in a neighborhood of $E(\bar{\mathbf{x}})$. See Shao and Tu 1995, Theorem 2.1.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics 16*, 152.164.

Fay, R. E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 495-500.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 212-218.

Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 732-736.

Fuller, W. (1998). Replicate variance estimation for two-phase samples. *Statistica Sinica 8*, 1153-1164.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics 6*, 223-240.

Kalton, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society A, 142, Part 2*, 210-222.

Rao, J. N. K., and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association 83*, 231-241.

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

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

Appendix – Unbiasedness of Totals Variance Estimator

In this appendix a proof that the variance estimator $v\left(\hat{Y}\right)$ in Section 3 is an unbiased estimator of the variance $Var\left(\hat{Y}\right)$ is given. We can write the variance as follows:

$$Var\left(\hat{Y}\right)=\sum_{t=1}^{11} Var\left(\hat{Y}_t\right)$$

$$\text{with}\quad Var\left(\hat{Y}_t\right)=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}$$
$$+\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{tis}\right)}{\pi_{ti}f_{tis}^2}S_{tisy}^2\quad \text{for } t=1,...,10,$$
$$\text{and}\quad Var\left(\hat{Y}_t\right)=\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{tis}\right)}{f_{tis}^2}S_{tisy}^2\quad \text{for } t=11.$$

Re-writing the variance estimator in a corresponding way, we obtain the following:

$$v\left(\hat{Y}\right)=\sum_{t=1}^{11} v\left(\hat{Y}_t\right)$$

$$\text{with}\quad v\left(\hat{Y}_t\right)=\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)}{\pi_{ti,tj}}\left\{\left(\frac{\hat{Y}_{ti}}{\pi_{ti}}-\frac{\hat{Y}_{tj}}{\pi_{tj}}\right)^2\right\}$$
$$+\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{tis}\right)}{\pi_{ti}f_{tis}^2}s_{tisy}^2\quad \text{for } t=1,...,10,$$
$$\text{and}\quad v\left(\hat{Y}_t\right)=\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{tis}\right)}{f_{tis}^2}s_{tisy}^2\quad \text{for } t=11.$$

As a first step, it is easy to show that $v\left(\hat{Y}_{11}\right)$ is an unbiased estimator of $Var\left(\hat{Y}_{11}\right)$. The remaining task is to show that each $v\left(\hat{Y}_t\right)$ (t=1,…,10) is an unbiased estimator of each $Var\left(\hat{Y}_t\right)$ (t=1,…,10) respectively. Showing this comprises the remainder of this appendix.

A preliminary result is as follows:

$$\sum_{j\neq i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)=\pi_{ti}\left(1-\pi_{ti}\right)$$

We can separate out the terms of $v\left(\hat{Y}_t\right)$ as follows:

$$v\left(\hat{Y}_t\right)=YG_t+WC_t=\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)}{\pi_{ti,tj}}\left(\frac{\hat{Y}_{ti}}{\pi_{ti}}-\frac{\hat{Y}_{tj}}{\pi_{tj}}\right)^2$$
$$+\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}f_{ti}^2}s_{tisy}^2$$

The first part of finding the expectation of $v\left(\hat{Y}_t\right)$ will be to find $E(YG_t)=E\{E(YG_t|S_t)\}$, where S_t indicates the realized PSU sample in the superstratum. Starting with the conditional expectation:

$$E\left(YG_t \mid S_t\right)=\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)}{\pi_{ti,tj}}\ E\left\{\left(\frac{\hat{Y}_{ti}}{\pi_{ti}}-\frac{\hat{Y}_{tj}}{\pi_{tj}}\right)^2 \mid S_t\right\}=$$

$$=\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)}{\pi_{ti,tj}}\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2+\frac{Var\left(\hat{Y}_{ti} \mid S_t\right)}{\pi_{ti}^2}=\right.$$

$$\left.+\frac{Var\left(\hat{Y}_{tj} \mid S_t\right)}{\pi_{tj}^2}\right\}$$

$$=\sum_{i=1}^{n_t}\sum_{j>i}^{n_t}\frac{\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)}{\pi_{ti,tj}}\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right.$$

$$\left.+\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}^2 f_{ti}^2}S_{tisy}^2+\sum_{s=1}^{S_{tj}}\frac{m_{tjs}\left(1-f_{tj}\right)}{\pi_{tj}^2 f_{tj}^2}S_{tjsy}^2\right\}$$

Taking then the expectation over all possible PSU samples S_t in the superstratum:

$$E\left(YG_t\right)=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right.$$

$$\left.+\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}^2 f_{ti}^2}S_{tisy}^2+\sum_{s=1}^{S_{tj}}\frac{m_{tjs}\left(1-f_{tj}\right)}{\pi_{tj}^2 f_{tj}^2}S_{tjsy}^2\right\}=$$

$$=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}$$

$$+\sum_{i=1}^{N_t}\sum_{j\neq i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}^2 f_{ti}^2}S_{tisy}^2\right\}=$$

$$=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}$$

$$+\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}^2 f_{ti}^2}S_{tisy}^2\left\{\sum_{j\neq i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\right\}=$$

$$=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}$$

$$+\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti}^2 f_{ti}^2}S_{tisy}^2\left(\pi_{ti}\left(1-\pi_{ti}\right)\right)=$$

$$=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)\ \left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}$$

$$+\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti} f_{ti}^2}S_{tisy}^2\left(1-\pi_{ti}\right)$$

For the second term of the variance (WC_t), the expectation is as follows:

$$E\left(WC_t\right)=E\left(\sum_{i=1}^{n_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti} f_{ti}^2}s_{tisy}^2\right)$$

$$=\sum_{i=1}^{N_t}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{f_{ti}^2}S_{tisy}^2$$

From these two expectations then

$$E\left(v\left(\hat{Y}_t\right)\right)=\ E\left(YG_t+WC_t\right)=\sum_{i=1}^{N_t}\sum_{j>i}^{N_t}\left(\pi_{ti}\pi_{tj}-\pi_{ti,tj}\right)$$

$$\left\{\left(\frac{Y_{ti}}{\pi_{ti}}-\frac{Y_{tj}}{\pi_{tj}}\right)^2\right\}+\sum_{i=1}^{N_{ss}}\sum_{s=1}^{S_{ti}}\frac{m_{tis}\left(1-f_{ti}\right)}{\pi_{ti} f_{ti}^2}S_{tisy}^2$$

which is equal to $Var\left(\hat{Y}_t\right)$, as desired.