DOUBLE SAMPLING IN A MULTI-STAGE DESIGN

David R. Judkins, Westat; Michael Hidiroglou, Office for National Statistics
David R. Judkins, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Controlled Selection; Variance Estimation; Rao-Sampford Selection

Abstract: Consider a multi-stage survey with unequal probabilities at each stage and rich information about last-stage units (LSUs) within sample penultimate clusters. One choice of sampling method for selecting the sample of LSUs under such a scenario would be to use stratified systematic PPS selection. The measure of size for the PPS procedure would be defined in terms of cumulative probabilities of selection across prior stages, but the strata and sort variables would be defined in terms of the information about the LSUs without reference to higher stage sample structure. This method is particularly simple to implement and effective at reducing variances on variables correlated with the strata and sort variables. However, it is well known that this procedure renders unbiased or even consistent estimation of variances impossible; hence *ad hoc* variance estimators must be employed. Some practitioners in the survey field have viewed this as an acceptable compromise between the goals of small variances and measurable precision. We propose an alternative procedure that satisfies constraints on sample sizes by last-stage stratum, keeps variances small, and admits unbiased variance estimates for linear statistics and consistent variance estimates for nonlinear statistics.

1. Introduction

Consider a multi-stage design where each stage is selected with unequal probabilities and without replacement. Suppose that the scheme is such that there is rich information available about the last-stage units (LSUs) within each penultimate cluster. Such information might be readily available on a frame or might be the product of new information gathering. The challenge is how to use this information to reduce sampling variances while still being able to estimate design-based variances consistently. In the framework of a single stage, the solution would be deep stratification and Neyman allocation or double sampling, also due to Neyman (1938). However, the theory for multi-stage sampling has not been well developed. Various schemes have been developed including controlled selection (Goodman and Kish, 1950; Causey, Cox, and Ernst, 1985) and various types of systematic PPS selection. Although these schemes are effective in reducing variances, good variance estimators have never been developed for them. (For an innovative approach on systematic PPS sampling, see Kaufman, 1999. For a more thorough description of alternatives, see Wolter, 1985.) We offer an approach that can realize much of the same gains in true variances while not sacrificing variance estimation.

Estimation and variance estimation follow the general framework for double sampling of Chapter 9 of Särndal, Swensson, and Wretman (1992). Binder et al. (2000) considered variance estimation for the generalized regression estimator in a two-phase context when the first-phase sample is re-stratified using information gathered from the first-phase sample. They provided simple computational expressions for variance estimation for the double expansion estimator and the reweighted expansion estimator of Kott and Stukel (1997). Bérard, Brodeur, and St. Pierre (1999) used this design and associated ratio estimators to compute estimated variances for data from the Canadian Retail Commodity Survey, which used this two-phase stratification. Hidiroglou and Rao (2003) recently developed a generalized procedure for the Yates-Grundy-Sen type variance estimator when sample sizes are fixed at both phases.

The paper is structured as follows. In Section 2, we develop a new strategy for variance estimation that can be applied to a very wide range of stratified two-stage designs where second-stage sampling need not be independent across sampled first-stage units. Such flexibility enables a variety of strategies for using the rich auxiliary information about last-stage units. The referenced work of Binder and coauthors are obtained as special cases of this new strategy. Section 3 outlines a specific stratified two-stage design that does use rich information about LSUs and that could be used in conjunction with the new strategy for variance estimation. Section 4 offers a brief summary and direction for future research.

1.1 Estimation and Variance Estimation for a General Stratified Two-Stage Sampling Scheme

Let the universe of PSUs be denoted as U. This universe is stratified into H strata U_h, and a sample of m_h PSUs is selected from M_h PSUs using an arbitrary sampling scheme. The second-stage units, within these the resulting overall sample, are re-stratified into another set of L strata. A set of second-stage units is further selected within the second-stage strata. We assume that the first-phase sample is stratified, without replacement, such that there are no zero joint selection probabilities: the second phase sample of second-stage units is also selected in such a way that all conditional joint probabilities of selection are positive. Such a set of assumptions would, for example, be reasonable if the Durbin-Brewer selection method were used at the first-stage and the Rao-Sampford method at the second-stage/phase. If only one PSU were selected per first-stage stratum, then it would be necessary to collapse strata and possibly develop some appropriate joint probabilities of selection as done by Shapiro and Bateman (1978).

To put the problem in algebraic context, let $h = 1, \ldots, H$ index the first-stage strata. Let M_h be the total number of PSUs and m_h be the number of sample PSUs in the h-th stratum. Let F_h denote the corresponding set of sample PSUs, and

$F = \bigcup_{h=1}^{H} F_h$. Let L be the total number of second-stage/phase strata. Let S_{hic}, $c = 1,\ldots,L$, be the set of second-stage units listed within second-stage stratum c and the i-th sample PSU in the h-th first-stage stratum, and let s_{hic} be the subset of those that were selected at the second phase. Let y_{hicj} be the value of some variable y for the indexed unit (either within the sample or the population, depending on summation limits), and let Y_{hi} be the total for y in the indexed PSU (either within the first-stage sample or the population, depending on summation limits).

Let π_{1hi} be the probability of selection at the first-stage for PSU i within first-stage stratum h; $\pi_{1hii'}$ be the joint probability of selection at the first-stage for PSUs i and i' within the first-stage stratum h; and $d_{1hii'} = \dfrac{\pi_{1hi}\pi_{1hi'} - \pi_{1hii'}}{\pi_{1hii'}}$. Similarly, let π_{2hicj} be the conditional probability of selection at the second-stage for the j-th listed or sampled unit within PSU i within first-stage stratum h; $\pi_{2hicjj'}$ be the joint probability of selection for two second-stage units in the same PSU of the same first-stage stratum and same second-stage stratum; $\pi_{2hii'cjj'}$ be the joint probability of selection for two second-stage units in different PSUs of the same first-stage stratum and same second-stage stratum; $\pi_{2hh'ii'cjj'}$ be the joint probability of selection for two second-stage units in different PSUs in different first-stage strata but same second-stage stratum.

Defining $\theta_{2hicjj'} = \dfrac{\pi_{2hicjj'}}{\pi_{2hicj}\pi_{2hicj'}} - 1$, $\theta_{2hii'cjj'} = \dfrac{\pi_{2hii'cjj'}}{\pi_{2hicj}\pi_{2hi'cj'}} - 1$, and $\theta_{2hh'ii'cjj'} = \dfrac{\pi_{2hh'ii'cjj'}}{\pi_{2hicj}\pi_{2h'i'cj'}} - 1$, the double-expansion or π^* estimator of Särndal et al. (1992, Chap. 9) for a population total on some variable y can be expressed as:

$$\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{c=1}^{L}\sum_{j\in s_{hic}} \frac{y_{hicj}}{\pi_{1hi}\pi_{2hicj}}.$$

The corresponding population variance of $\hat{Y}$ can be obtained via the following well-known identity:

$$\mathrm{Var}\left(\hat{Y}\right) = \mathrm{Var}\left[\mathrm{E}\left(\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{c=1}^{L}\sum_{j\in s_{hic}} \frac{y_{hicj}}{\pi_{1hi}\pi_{2hicj}}\middle| F\right)\right] + \mathrm{E}\left[\mathrm{Var}\left(\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{c=1}^{L}\sum_{j\in s_{hic}} \frac{y_{hicj}}{\pi_{1hi}\pi_{2hicj}}\middle| F\right)\right]$$

$$= \mathrm{Var}\left(\sum_{h=1}^{H}\sum_{i=1}^{m_h}\frac{Y_{hi}}{\pi_{1hi}}\right) + \mathrm{E}\left[\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{c=1}^{L}\sum_{j\in S_{hic}} \frac{y_{hicj}^2}{\pi_{1hi}^2}\left(\frac{1}{\pi_{2hicj}} - 1\right)\right]$$
$$+ 2\mathrm{E}\left[\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{c=1}^{L}\sum_{j\in S_{hic}}\sum_{\substack{j'\in S_{hic}\\ j'>j}} \theta_{2hicjj'}\frac{y_{hicj}y_{hicj'}}{\pi_{1hi}^2}\right]$$
$$+ 2\mathrm{E}\left[\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}\sum_{c=1}^{L}\sum_{j\in S_{hic}}\sum_{j'\in S_{hi'c}} \theta_{2hii'cjj'}\frac{y_{hicj}}{\pi_{1hi}}\frac{y_{hi'cj'}}{\pi_{1hi'}}\right]$$
$$+ 2\mathrm{E}\left[\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{h'=h+1}^{H}\sum_{i'=1}^{m_{h'}}\sum_{c=1}^{L}\sum_{j\in S_{hic}}\sum_{j'\in S_{h'i'c}} \theta_{2hh'ii'cjj'}\frac{y_{hicj}}{\pi_{1hi}}\frac{y_{h'i'cj'}}{\pi_{1h'i'}}\right]$$

Denote each line of the above equation respectively as V_1, T_2, T_3, T_4 and T_5: that is, the population variance $\mathrm{Var}\left(\hat{Y}\right)$ can be expressed as $V_1 + T_2 + T_3 + T_4 + T_5$, where V_1 is the ordinary between-PSU variance, $T_2 + T_3$ is the within-PSU variance, and $T_4 + T_5$ is an irregular adjustment to between-PSU variance, including cross-terms across first-stage strata. The expressions T_3, T_4, and T_5 may all be negative. Further simplification of these terms is difficult because the second-stage probabilities are random variables, depending on F. Also note that the familiar Yates-Grundy form of the Horvitz-Thompson estimator cannot be used because the sample sizes within second-stage strata are not fixed.

We now construct a variance estimator for the population variance $\mathrm{Var}\left(\hat{Y}\right)$. Following Hidiroglou and Rao (2003), we choose a Yates-Grundy-Sen type estimator V_1 as an initial candidate, where

$$\hat{V}_1 = \sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h} d_{1hii'}\left(\frac{\sum_{c=1}^{L}\sum_{j\in s_{hic}}\frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{c=1}^{L}\sum_{j\in s_{hi'c}}\frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}}\right)^2.$$

We derive the expected value of this estimator as a step toward obtaining an unbiased estimator of $\mathrm{Var}\left(\hat{Y}\right)$.

$$\mathrm{E}\left(\hat{V}_1\right) = \mathrm{E}\left\{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h} d_{1hii'} \times \mathrm{E}\left[\left(\frac{\sum_{c=1}^{L}\sum_{j\in s_{hic}}\frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{c=1}^{L}\sum_{j\in s_{hi'c}}\frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}}\right)^2\middle| F\right]\right\}$$

$$= \mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \left\{ \mathrm{Var}\left[\left(\frac{\sum_{c=1}^{L} \sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{c=1}^{L} \sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \right) \middle| F \right] + \left(\frac{Y_{hi}}{\pi_{1hi}} - \frac{Y_{hi'}}{\pi_{1hi'}} \right)^2 \right\} \right\}$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{i'=i+1}^{M_h} d_{1hii'} \pi_{1hii'} \left(\frac{Y_{hi}}{\pi_{1hi}} - \frac{Y_{hi'}}{\pi_{1hi'}} \right)^2 + \mathrm{E}\left[\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \mathrm{Var}\left(\frac{\sum_{c=1}^{L} \sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{c=1}^{L} \sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \middle| F \right) \right]$$

$$= V_1 + \mathrm{E}\left[\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \mathrm{Var}\left(\sum_{c=1}^{L} \left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \right) \middle| F \right) \right]$$

Since the second-phase sampling is conditionally independent across second-phase strata, the variance operator can be brought within the summation over second-phase strata so that

$$\mathrm{E}\left(\hat{V}_1\right) = V_1 + \mathrm{E}\left[\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \sum_{c=1}^{L} \mathrm{Var}\left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} - \frac{\sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \middle| F \right) \right]$$

$$= V_1 + \mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \sum_{c=1}^{L} \left[\mathrm{Var}\left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} \middle| F \right) + \mathrm{Var}\left(\frac{\sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \middle| F \right) - 2\mathrm{Cov}\left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}}, \frac{\sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \middle| F \right) \right] \right\}$$

$$= V_1 + \mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \times \sum_{c=1}^{L} \left[2\mathrm{Var}\left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}} \middle| F \right) + -2\mathrm{Cov}\left(\frac{\sum_{j \in s_{hic}} \frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}}, \frac{\sum_{j \in s_{hi'c}} \frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}} \middle| F \right) \right] \right\}$$

We now use Horvitz and Thompson's original formula (Horvitz and Thompson, 1952) to rewrite the conditional variance in the line above, and we use Lemmas 1 and 2 from the appendix to rewrite the conditional covariance. This gives us

$$\begin{aligned} \mathrm{E}\left(\hat{V}_1\right) = V_1 &+ 2\mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} d_{1hii'} \sum_{c=1}^{L} \sum_{j \in S_{hic}} \frac{y_{hicj}^2}{\pi_{1hi}^2} \left(\frac{1}{\pi_{2hicj}} - 1 \right) \right\} \\ &+ 2\mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} \frac{d_{1hii'}}{\pi_{1hi}^2} \times \sum_{c=1}^{L} \sum_{j \in S_{hic}} \sum_{\substack{j' \in S_{hic} \\ j' \neq j}} \theta_{2hicjj'} y_{hicj} y_{hicj'} \right\} \\ &- 2\mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} \frac{d_{1hii'}}{\pi_{1hi} \pi_{1hi'}} \times \sum_{c=1}^{L} \sum_{j \in S_{hic}} \sum_{j' \in S_{hi'c}} \theta_{2hii'cjj'} y_{hicj} y_{hi'cj'} \right\} \end{aligned}$$

Note that in the last step, the index i' appears only in the $d_{1hii'}$ factor of the second and third terms. This allows a conditioning strategy to pull this factor out. A similar strategy does not work on the fourth term because the index i' appears in several factors. Hence,

$$\begin{aligned} \mathrm{E}\left(\hat{V}_1\right) = V_1 + T_2 - \mathrm{E}&\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \frac{1}{\pi_{1hi}} \sum_{c=1}^{L} \sum_{j \in S_{hic}} y_{hicj}^2 \left(\frac{1}{\pi_{2hicj}} - 1 \right) \right\} + T_3 \\ &- 2\mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \frac{1}{\pi_{1hi}} \sum_{c=1}^{L} \sum_{j \in S_{hic}} \sum_{\substack{j' \in S_{hic} \\ j' > j}} \theta_{2hicjj'} y_{hicj} y_{hicj'} \right\} \\ &- 2\mathrm{E}\left\{ \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{i'=i+1}^{m_h} \frac{d_{1hii'}}{\pi_{1hi} \pi_{1hi'}} \times \sum_{c=1}^{L} \sum_{j \in S_{hic}} \sum_{j' \in S_{hi'c}} \theta_{2hii'cjj'} y_{hicj} y_{hi'cj'} \right\} \end{aligned}$$

We write $\mathrm{E}\left(\hat{V}_1\right)$ as $V_1+\mathrm{T}_2-B_2+T_3-B_3-B_4$. The V's have been previously defined and the B's are the last two terms of the above equation.

Given unbiased estimators for the three biases B_2, B_3, and B_4 as well as unbiased estimators for the terms T_4 and T_5, we can construct an unbiased variance estimator as $\hat{V}\left(\hat{Y}\right)=\hat{V}_1+\hat{B}_2+\hat{B}_3+\hat{B}_4+\hat{T}_4+\hat{T}_5$. The natural choices for these estimators are

$$\hat{B}_2=\sum_{h=1}^{H}\sum_{i=1}^{m_h}\frac{1}{\pi_{1hi}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\frac{y_{hicj}^2}{\pi_{2hicj}}\left(\frac{1}{\pi_{2hicj}}-1\right)$$

$$\hat{B}_3=2\sum_{h=1}^{H}\sum_{i=1}^{m_h}\frac{1}{\pi_{1hi}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{\substack{j'\in s_{hic}\\ j'>j}}\frac{\theta_{2hicjj'}}{\pi_{2hicjj'}}y_{hicj}y_{hicj'}$$

$$\hat{B}_4=2\sum_{\mathrm{h}=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}\frac{d_{1hii'}}{\pi_{1hi}\pi_{1hi'}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{j'\in s_{hi'c}}\frac{\theta_{2hii'cjj'}}{\pi_{2hii'cjj'}}y_{hicj}y_{hi'cj'}$$

$$\hat{T}_4=2\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{j'\in s_{hi'c}}\frac{\theta_{2hii'cjj'}}{\pi_{2hii'cjj'}}\frac{y_{hicj}}{\pi_{1hi}}\frac{y_{hi'cj'}}{\pi_{1hi'}}$$

$$\hat{T}_5=2\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{h'=h+1}^{H}\sum_{i'=1}^{m_{h'}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{j'\in s_{h'i'c}}\frac{\theta_{2hh'ii'cjj'}}{\pi_{2hh'ii'cjj'}}\frac{y_{hicj}}{\pi_{1hi}}\frac{y_{h'i'cj'}}{\pi_{1h'i'}}$$

Adding these terms, and simplifying, we obtain the estimated variance for $\hat{Y}$ as:

$$\begin{aligned}\hat{V}(\hat{Y})=&\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}d_{1hii'}\left(\frac{\sum_{c=1}^{L}\sum_{j\in s_{hic}}\frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}}-\frac{\sum_{c=1}^{L}\sum_{j\in s_{hi'c}}\frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}}\right)^2\\&+\sum_{h=1}^{H}\sum_{i=1}^{m_h}\frac{1}{\pi_{1hi}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\left[\frac{y_{hicj}^2}{\pi_{2hicj}}\left(\frac{1}{\pi_{2hicj}}-1\right)\right.\\&\left.+2\sum_{\substack{j'\in s_{hic}\\ j'>j}}\frac{\theta_{2hicjj'}}{\pi_{2hicjj'}}y_{hicj}y_{hicj'}\right]\\&+2\sum_{\mathrm{h}=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}\frac{1}{\pi_{1hii'}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{j'\in s_{hi'c}}\frac{\theta_{2hii'cjj'}}{\pi_{2hii'cjj'}}y_{hicj}y_{hi'cj'}\\&+2\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{h'=h+1}^{H}\sum_{i'=1}^{m_{h'}}\sum_{c=1}^{L}\sum_{j\in s_{hic}}\sum_{j'\in s_{h'i'c}}\frac{\theta_{2hh'ii'cjj'}}{\pi_{2hh'ii'cjj'}}\frac{y_{hicj}}{\pi_{1hi}}\frac{y_{h'i'cj'}}{\pi_{1h'i'}}\end{aligned}$$

This construction completes the proof of Theorem 1.

Theorem 1: Consider a two-stage design with stratified without replacement sampling at each stage and fixed sample sizes at the first stage. Assume that first-stage sampling is independent across first-stage strata and the second-stage sampling is independent across second-stage strata. If all pairs of first-stage units have nonzero joint probabilities of selection, and if all pairs of second-stage units also have nonzero joint probabilities of selection, then $\hat{V}(\hat{Y})$ is an unbiased estimator of the variance of the double expansion estimator of population totals.

Remark 1: There is no requirement for fixed sample sizes for the second-stage strata. Furthermore, note that there is no requirement for second-stage sampling to be independent across PSUs, nor even across first-stage strata. This makes the estimator ideal for use in double-sampling applications.

Remark 2: An alternative expression for the variance estimator in Theorem 1 can be obtained by re-indexing the final stage sample by a single subscript and defining $\pi_{\ell\ell'}$ to be the overall joint probability of two final-stage units. Then

$$\begin{aligned}\hat{V}(\hat{Y})=&\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{i'=i+1}^{m_h}d_{1hii'}\left(\frac{\sum_{c=1}^{L}\sum_{j\in s_{hic}}\frac{y_{hicj}}{\pi_{2hicj}}}{\pi_{1hi}}-\frac{\sum_{c=1}^{L}\sum_{j\in s_{hi'c}}\frac{y_{hi'cj}}{\pi_{2hi'cj}}}{\pi_{1hi'}}\right)^2\\&+\sum_{\ell}\frac{y_\ell^2}{\pi_{1\ell}\pi_{2\ell}}\left(\frac{1}{\pi_{2\ell}}-1\right)+\sum_{\ell}\sum_{\ell'\neq\ell}\frac{\theta_{2\ell\ell'}}{\pi_{\ell\ell'}}y_\ell y_{\ell'}\end{aligned}$$

Corollary 1: Consider a two-stage design where each stage is SWSWOR replacement and there is no stratification. Let $f_1=m/M$ be the first-stage sampling fraction, $N_{|F}$ be the number of second-stage units listed in the selected PSUs, and $f_{2|F}=n/N_{|F}$ be the second-stage sampling fraction given the selected set of n units within the $N_{|F}$ units listed within the selected set of PSUs. Then the variance on the double expansion estimator of a population total is

$$\begin{aligned}V(\hat{Y})=&\frac{M}{m}\frac{1-f_1}{M-1}\sum_{i=1}^{M}\sum_{i'=i+1}^{M}\left(Y_i-Y_{i'}\right)^2\\&+\frac{M^2}{m^2}\mathrm{E}\left[\sum_{i=1}^{m}\frac{N_{|F}\left(1-f_{2|F}\right)}{n}\times\right.\\&\left.\sum_{j\in S_i}\left[{y_{ij}}^2-\frac{2}{(N_{|F}-1)}\sum_{\substack{j'\in S_i\\ j'>j}}y_{ij}y_{ij'}\right]\right]\\&+\frac{2M^2}{m^2}\mathrm{E}\left[\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\frac{N_{|F}\left(1-f_{2|F}\right)}{n(N_{|F}-1)}\sum_{j\in S_i}\sum_{j'\in S_{i'}}y_{ij}y_{i'j'}\right]\end{aligned}$$

and an unbiased estimator of this variance is

$$\hat{V}(\hat{Y}) = \frac{M^2}{m^2}\frac{1-f_1}{m-1}\frac{N_{|F}^2}{n^2}\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\left(\sum_{j\in s_i} y_{ij} - \sum_{j\in s_{i'}} y_{i'j}\right)^2$$
$$+\frac{M}{m}\frac{N_{|F}^2}{n(n-1)}\left(1-f_{2|F}\right)\sum_{i=1}^{m}\sum_{j\in s_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$$
$$+\frac{M}{m}\frac{N_{|F}^2}{n(n-1)}\left(1-f_{2|F}\right)\sum_{i=1}^{m}\left(1-\frac{n_i}{n}\right)n_i\bar{y}_{i.}^2$$
$$-2\frac{M}{m}\frac{M-1}{m-1}\frac{N_{|F}^2}{n^2}\frac{\left(1-f_{2|F}\right)}{(n-1)}\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\sum_{j\in s_i}\sum_{j'\in s_{i'}} y_{ij}y_{i'j'}$$

where $\bar{y}_{i.} = \sum_{j\in s_i} y_{ij}/n_i$ and n_i is the number of sampled second-stage units within PSU i.

Corollary 2: Consider a two-stage design where the first stage is SWSWOR replacement and the second stage is stratified SRSWOR. Let n_c be the overall number of selected second-stage units within second-stage stratum c and n_{ic} be the corresponding number of selected second-stage units within that stratum and PSU i. Then the variance on the double expansion estimator of a population total is

$$V(\hat{Y}) = \frac{M}{m}\frac{1-f_1}{M-1}\sum_{i=1}^{M}\sum_{i'=i+1}^{M}\left(Y_i - Y_{i'}\right)^2$$
$$+\frac{M^2}{m^2}\mathrm{E}\left[\sum_{c=1}^{L}\sum_{i=1}^{m}\frac{N_{c|F}\left(1-f_{2c|F}\right)}{n_c}\times\right.$$
$$\left.\sum_{j\in S_{ic}}\left[y_{icj}^2 - \frac{2}{(N_{c|F}-1)}\sum_{\substack{j'\in S_{ic}\\ j'>j}} y_{icj}y_{icj'}\right]\right]$$
$$+\frac{2M^2}{m^2}\mathrm{E}\left[\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\frac{N_{c|F}\left(1-f_{2c|F}\right)}{n_c(N_{c|F}-1)}\sum_{j\in S_{ic}}\sum_{j'\in S_{i'c}} y_{icj}y_{i'cj'}\right]$$

and an unbiased estimator of this variance is

$$\hat{V}(\hat{Y}) = \frac{M^2}{m^2}\frac{1-f_1}{m-1}\frac{N_{c|F}^2}{n_c^2}\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\left(\sum_{c=1}^{L}\sum_{j\in s_{ic}} y_{icj} - \sum_{c=1}^{L}\sum_{j\in s_{i'c}} y_{i'cj}\right)^2 +$$
$$+\frac{M}{m}\sum_{i=1}^{m}\sum_{c=1}^{L}\frac{N_{c|F}^2}{n_c n_c - 1}\left(1-f_{2c|F}\right)\times$$
$$\left[\sum_{j\in s_{ic}}\left(y_{icj}-\bar{y}_{ic.}\right)^2 + \bar{y}_{ic.}^2 n_{ci}\left(\left(1-\frac{n_{ci}}{n_c}\right)\right)\right]+$$
$$-2\frac{M}{m}\frac{M-1}{m-1}\sum_{i=1}^{m}\sum_{i'=i+1}^{m}\sum_{c=1}^{L}\frac{N_{c|F}^2}{n_c^2}\frac{\left(1-f_{2c|F}\right)}{(n_c-1)}\times$$
$$\sum_{j\in s_{ic}}\sum_{j'\in s_{i'c}} y_{icj}y_{i'cj'}$$

3. A Specific Stratified Two-Stage Design

Having worked out a variance estimation strategy for a wide range of stratified two-stage designs where the second-stage sampling need not be independent across sample PSUs, we now lay out a specific example of a sampling scheme that should result in good utilization of rich LSU information. We first summarize the steps involved in the scheme and then discuss each in more detail:

1. Stratify the list of all last-stage units from sample penultimate-stage units as deeply as sensible, using all auxiliary information except possibly detailed size. (If there are sample requirements for last-stage domains, it might be as simple as using these domains as strata.);
2. Define a measure of size for each listed last-stage unit, taking into account probabilities of selection at prior stages so as to achieve a self-weighting sample within each last-stage stratum;
3. Identify certainty selections as those where the measure of size is greater than 1/n-th of the cumulative measure of size for the stratum, where n is the desired sample size for the stratum. Adjust the desired sample size from among the smaller units in the stratum accordingly;
4. Use the Rao-Sampford selection procedure (Rao 1965; Sampford, 1967) to select the required sample from the noncertainties within each stratum. This method will select the sample to have the exact probabilities of selection desired, the exact sample size desired, and no unit selected more than once; and;
5. Compute weights appropriate for double expansion estimators (Särndal et al., 1992). Use the new variance estimator developed earlier in the paper.

3.1 Deep Explicit Stratification

Other than simplicity, the main benefit of systematic PPS selection is the possibility to reduce variances through implicit stratification on more characteristics than are used to form the explicit strata. However, there are alternatives to simply using PPSWOR sampling within the explicit strata. If one has more auxiliary information than is reflected in the explicit strata and the measure of size, one can always use that extra information to stratify more deeply. Jewett and Judkins (1988), later improved by Ludington (1992), developed an algorithm for automatic optimal stratification that minimizes the variance in stratum cumulative measures of size while also maximizing between-stratum variance on the auxiliary information. Even without automatic software, deeper stratification can be done manually fairly easily if one is satisfied with Cartesian products of a few variables.

3.2 Measures of Size

We describe the operations for a two-stage design, but the generalization to three or more stages is simple. The notation is mostly the same as in Section 2. Let Z_{hicj} be a measure of size for the j-th listed or sampled unit within PSU i within the first-

stage stratum *h*. Assume that Z_{hicj} is independent of the sample design (e.g., employee count of an establishment). Let n_c be the desired national sample size for the *c*-th second-stage stratum, and let N_{hic} be the number of second-stage units listed for the stratum within the *i*-th sample PSU within first-stage stratum *h*. Then the second-phase probabilities should be set as

$$\pi_{2hicj} = \frac{n_c}{\pi_{1hi}} \frac{Z_{hicj}}{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \frac{1}{\pi_{1hi}} \sum_{k=1}^{N_{hic}} Z_{hick}}.$$

It can be easily verified that $\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{N_{hic}} \pi_{2hicj} = n_c$, as desired. These probabilities can be achieved by using the simpler measure of size

$$T_{hicj} = \frac{Z_{hicj}}{\pi_{1hi}},$$

the same as typically used for systematic PPS selection.

3.3 Conditional Certainties

Some second-stage units with either large size or small first-stage probabilities may have to be selected with certainty. The cutoff is

$$T_{hicj} > \frac{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \frac{1}{\pi_{1hi}} \sum_{k=1}^{N_{hic}} Z_{hick}}{n_c}.$$

In practice, we have often seen people select units with measures of size greater than 75 percent of the maximum certainty threshold selected with certainty in order to reduce problems with unstable variance estimates. However, that may be mainly a concern at the first stage of selection and need not be followed at the second stage.

3.4 Rao-Sampford Selection

This method due to Rao (1965) and Sampford (1967) is described in Brewer and Hanif (1983) among other places, although there are several typographic errors in the relevant section of this text that makes reference to the Sampford paper important. Calculation of the joint probabilities induced by this method used to be difficult for $n_c > 2$, but Mecatti, Haziza, and Rao (2004) have developed software that solves this problem quickly on modern computers. Another solution is to use an approximation due to Asok and Sukhatme (1976).

We recommend the Rao-Sampford sampling method based on several considerations. First, the variance estimator suggested below requires that all joint probabilities of selection at both phases of sampling be strictly positive. This is achieved by using stratified Rao-Sampford selection. Second, Brewer and Hanif (1983) found that calculation of joint probabilities of selection was easiest with the Rao-Sampford procedure and this should be even easier with the new software from Mecatti et al. (2004). Third, Brewer and Hanif found that the stability of the Yates-Grundy variance estimator applied to samples selected with the Rao-Sampford method was near the lower limit that can be achieved with any sampling procedure. Fourth, Bayless and Rao (1970) found that for a fairly general set of superpopulations, the model expectation of the design-based variance does not depend on the joint probabilities of selection, so that any method is as a good as another for minimizing variances. (The model is only reasonable, however, within strata that have fully exploited all auxiliary information other than the size information.) However, there are many other reasonable choices for without replacement sampling. In addition to Brewer and Hanif (1983), see Chaudhuri and Vos (1988).

4. Summary and Future Research

We have developed a new general variance estimator for linear estimates from stratified two-stage designs where second-stage sampling is dependent across PSUs. It should be possible to extend this to nonlinear statistics through either linearization (e.g., by adding it as a design to SUDAAN) or replication (following the general framework of Fay, 1984, and 1989). We have also proposed a specific strategy for stratified two-stage sampling that makes extensive use of knowledge about second-stage units in the frame. Areas where this could be applied include clustered sampling of establishments from administrative lists and double sampling of second-stage units. With these innovations, we think there is no longer any compelling reason to use the common practice of using systematic PPS selection across PSUs to select second-stage samples.

5. References

Asok, C., and Sukhatme, B.V. (1976). On Sampford's Procedure of Unequal Probability Sampling Without Replacement. *Journal of the American Statistical Association*, **71**, 912-918.

Binder, D.A., Babyak, C. , Brodeur, M., Hidiroglou, M., and Jocelyn W. (2000). Variance Estimation for Two-Phase Stratified Sampling. *The Canadian Journal of Statistics*, **28**, No. 4, 751-764.

Bérard, H., Brodeur, M., and St. Pierre, M. (1999). The Retail Commodity Survey. *Proceedings of the Survey Methods Section,* Statistical Society of Canada, 141-146.

Bayless, D.L., and Rao, J. N. K. (1970). An Empirical Study of the Stabilities of Estimators and Variance Estimators in Unequal Probability Sampling ($n = 3$ or 4). *Journal of the American Statistical Association*, **65**, 1645-1667.

Brewer, K.R.W., and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Causey, B.D., Cox, L.H., and Ernst, L.R. (1985). Applications of Transportation Theory to Statistical Problems. *Journal of the American Statistical Association*, **80**, 903-909.

Chaudhuri, A., and Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. New York: Elsevier/North-Holland.

Fay, R.E. (1984). Some Properties of Estimators of Variance Based on Replication Methods. *Proceedings of the Section on Survey Research Methods of the 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 of the American Statistical Association*, 212-217.

Goodman, R., and Kish, L. (1950). Controlled Selection – A Technique in Probability Sampling. *Journal of the American Statistical Association*, **45**, 350-372.

Hidiroglou, M.A., and Rao, J.N.K. (2003). Variance Estimation in Two-phase Sampling. *Statistics Canada Methodology Symposium.*

Horvitz, D.G., and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

Jewett, R., and Judkins, D. (1988). Multivariate Stratification with Size Constraints. *SIAM Journal on Scientific and Statistical Computing*, **9**, 1091-1097.

Kaufman, S. (1999). Using the Bootstrap to Estimate the Variance From a Single Systematic PPS Sample. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 683-688.

Kott P.S., and Stukel D.M. (1997). Can the Jackknife be Used with a Two-Phase Sample? *Survey Methodology*, **23**, 81-89.

Ludington, P.W. (1992). Stratification of Primary Sampling Units for the Current Population Survey Using Computer Intensive Methods. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 752-757.

Mecatti, F., Haziza, D., and Rao, J.N.K. (2004). Comparison of Variance Estimators Under Rao-Sampford Method: A Simulation Study. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **33**, 101-116.

Rao, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling Without Replacement. *Journal of the Indian Statistical Association*, **3**, 173-180.

Sampford, M.R. (1967). On Sampling Without Replacement with Unequal Probabilities of Selection. *Biometrika*, **54**, 499-513.

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

Shapiro, G. M., and Bateman, D.V. (1978). A Better Alternative to the Collapsed Stratum Variance Estimator. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 451-456.

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

APPENDIX

Lemma 1. Covariance of Two Horvitz-Thompson Estimators

Consider the simple case of an arbitrary PPS single stage unstratified design. Let N be the total number of units in the frame. Let π_i be the probability of selection for the i-th unit in the frame, δ_i be a sample indicator flag for the unit, and x_i and y_i be two variates. Let X and Y be the respective sums of these variates over the frame. Then the covariance between the Horvitz-Thompson estimators for X and Y is

$$\mathrm{Cov}\left(\sum_{i=1}^{N}\frac{\delta_i}{\pi_i}x_i,\sum_{i=1}^{N}\frac{\delta_i}{\pi_i}y_i\right)=\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)x_i y_i+\sum_{i=1}^{N}\sum_{j\neq i}\theta_{ij}x_i y_j$$

where $\theta_{ij}=\dfrac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}$.

Lemma 2. Covariance of Two Horvitz-Thompson Estimators of Population Sizes for Disjoint Domains

Under the same conditions as Lemma 1, suppose that and δ_{Ai} is a binary indicator flag for membership in Domain A, that δ_{Bi} is a binary indicator flag for membership in Domain B and that $A\cap B=\varnothing$. Let X_A be the total of x on domain A and Y_B be the total of y on domain B. Then the covariance between the HT estimates of X_A and Y_B is

$$\mathrm{Cov}\left(\sum_{i\in A}\frac{\delta_i}{\pi_i}x_i,\sum_{i\in B}\frac{\delta_i}{\pi_i}y_i\right)=\sum_{i\in A}\sum_{j\in B}\theta_{ij}x_i y_i\,.$$