Pseudo–Empirical Likelihood Inference for Multiple Frame Surveys

J. N. K. Rao and Changbao Wu

This article presents a pseudo–empirical likelihood approach to inference for multiple-frame surveys. We establish a unified framework for point and interval estimation of finite population parameters, and show that inferences on the parameters of interest making effective use of different types of auxiliary population information can be conveniently carried out through the constrained maximization of the pseudo–empirical likelihood function. Confidence intervals are constructed using either the asymptotic χ^2 distribution of an adjusted pseudo–empirical likelihood ratio statistic or a bootstrap calibration method. Simulation results based on Statistics Canada's Family Expenditure Survey data show that the proposed methods perform well in finite samples for both point and interval estimation. In particular, a multiplicity-based pseudo–empirical likelihood method is proposed. This method is easily used for multiple-frame surveys with more than two frames and does not require complete frame membership information. The proposed pseudo–empirical likelihood ratio confidence intervals have a clear advantage over the conventional normal approximation–based intervals in estimating population proportions of rare items, a scenario that often motivates the use of multiple-frame surveys. All related computational problems can be handled using existing algorithms for pseudo–empirical likelihood methods with single-frame surveys.

KEY WORDS: Confidence intervals; Design effect; Dual-frame surveys; Multiplicity; Survey design; Unequal probability sampling.

1. INTRODUCTION

Multiple-frame surveys are widely used by large statistical agencies and business organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. The Canadian Community Health Survey (CCHS), a cross-sectional survey that collects information related to health status, health system utilization, and health determinants for the Canadian population, uses a three-frame sampling design. The primary sampling frame is the area frame initially designed for the Canadian Labour Force Survey (LFS). A Random Digit Dialing sampling frame and a list frame of residential telephone numbers are also used to increase the frame coverage of the target population. In multiple-frame surveys, two or more population frames are available, each of which can be incomplete, but together they are assumed to cover the entire target population. Independent probability samples, one from each frame, are taken, and the goal is to make inference on the overall population parameters of interest using the combined sample data. When the sample data are collected using two frames, the problem is referred to as a dual-frame survey.

Multiple-frame surveys have been studied by several authors, with primary focus on point estimation (see Hartley 1962, 1974; Fuller and Burmeister 1972; Bankier 1986; Kalton and Anderson 1986; Skinner 1991; Skinner and Rao 1996). Lohr and Rao (2000) studied variance estimation for dual-frame surveys using the jackknife method and examined the performance of the associated normal approximation–based confidence intervals through simulation. Lohr and Rao (2006) derived optimal linear estimators and pseudo–maximum likelihood estimators for the population total when samples were taken independently from multiple frames (more than two) using probability sampling designs. They also provided a short discussion on variance estimation and showed that the asymptotic variance of their proposed estimators has a very complex form. Confidence intervals under multiple-frame surveys have not yet been studied in detail, however. Moreover, the systematic use of known auxiliary population information has not been investigated.

Empirical likelihood method was first introduced by Owen (1988) as a nonparametric inference tool for independent and identically distributed observations. For single-frame complex survey data, Chen and Sitter (1999) proposed a pseudo–empirical likelihood approach and discussed point estimation in the context of using auxiliary population information. Pseudo–empirical likelihood ratio confidence intervals for single-frame surveys were studied by Wu and Rao (2006). In this article we present a pseudo–empirical likelihood (PEL) approach to inference from multiple-frame surveys. Our proposed approach addresses both point estimation and confidence intervals, and known auxiliary population information can be conveniently incorporated into inferences through constrained maximization of the PEL function. Confidence intervals for the population mean or the finite population distribution function can be constructed using either a χ^2 approximation to the adjusted PEL ratio statistics or a bootstrap calibration method. The χ^2 approximation requires calculation of design effects, which involves variance estimation. The bootstrap method, on the other hand, bypasses the need for variance estimation and is valid for single-stage sampling designs with small sampling fractions. Simulation studies show that the bootstrap method performs well even when the sampling fractions are not very small. Our proposed multiplicity-based PEL approach is particularly appealing because it does not require complete frame membership information and is very easy to use for multiple- (i.e., more than two) frame surveys. It also provides superior results on both point and interval estimation. Our proposed PEL ratio confidence intervals have a clear advantage over the conventional

J. N. K. Rao is Distinguished Research Professor, School of Mathematics and Statistics, Carleton University, Ottawa ON K1S 5B6, Canada (E-mail: *jrao@math.carleton.ca*). Changbao Wu is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1, Canada (E-mail: *cbwu@uwaterloo.ca*). This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada. We are grateful to the editor, an associate editor, and two referees for their many valuable suggestions.

Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Theory and Methods
DOI: 10.1198/jasa.2010.tm09534

normal approximation–based intervals in estimating population proportions of rare items, a scenario that often motivates the use of multiple-frame surveys. All required computational procedures for the proposed methods on multiple-frame surveys can be handled using existing procedures for single-frame surveys after suitable reformulation of the involved maximization problem.

In Section 2 we present PEL methods for dual-frame surveys based on poststratified samples. Generalization of this approach to multiple-frame surveys is possible, but is heavy and cumbersome in notation. The approach also requires complete frame membership information and correct partitioning of samples into domains defined by the multiple frames. In Section 3 we present a multiplicity-based PEL approach, which not only requires less detailed frame membership information, but also involves no additional difficulty or notational complexity in dealing with general multiple-frame surveys. We present a bootstrap calibration method for confidence intervals in Section 4. In Section 5 we report results from an extensive simulation study, using data from Statistics Canada's Family Expenditure Survey, on the finite-sample performances of our proposed methods compared with existing approaches. We provide some concluding remarks in Section 6. Proofs and regularity conditions are given in the Appendix.

2. PSEUDO–EMPIRICAL LIKELIHOOD INFERENCE FOR DUAL–FRAME SURVEYS

In this section we present PEL methods for dual-frame surveys based on poststratified samples. We first describe the basic setting for dual-frame surveys, following the classical works of Hartley (1962, 1974; see also Skinner and Rao 1996 and Lohr and Rao 2000, 2006). Let A and B denote two sampling frames. Both frames can be incomplete, but it is assumed that together they cover the entire finite population, $\mathcal{U}$. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ be the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c$, $b = \mathcal{A}^c \cap \mathcal{B}$, and $ab = \mathcal{A} \cap \mathcal{B}$, such that $\mathcal{U} = \mathcal{A} \cup \mathcal{B} = a \cup b \cup ab$. Note that $\mathcal{A}^c$ and $\mathcal{B}^c$ denote complement sets of $\mathcal{A}$ and $\mathcal{B}$. Let N, N_A, N_B, N_a, N_b, and N_{ab} be the number of population units in $\mathcal{U}$, $\mathcal{A}$, $\mathcal{B}$, a, b, and ab, respectively. It follows that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ab}$, and $N = N_a + N_b + N_{ab} = N_A + N_B - N_{ab}$. Let $\bar{Y}$, $\bar{Y}_a$, $\bar{Y}_b$, and $\bar{Y}_{ab}$ denote the population or domain means of the response variable y for $\mathcal{U}$, a, b, and ab, respectively. It follows that

$$\bar{Y} = \frac{N_a}{N}\bar{Y}_a + \frac{N_b}{N}\bar{Y}_b + \frac{N_{ab}}{N}\bar{Y}_{ab}. \tag{2.1}$$

The main objective is to make inference about $\bar{Y}$ using dual-frame samples as well as any available auxiliary population information. This problem depends crucially on what is known about N_A, N_B, and N_{ab}, however.

We consider three special cases:

1. N_A, N_B, and N_{ab} are all known
2. N_A and N_B are known but N_{ab} is unknown
3. N_A, N_B, and N_{ab} are all unknown.

Skinner and Rao (1996) addressed practical situations in which these three cases may arise. We focus on the first case and provide brief discussions on how to extend the proposed method to the second and third cases.

Case 1: N_A, N_B, and N_{ab} All Known

We first consider scenarios in which none of the two frames is complete and hence $N_a > 0$ and $N_b > 0$. The sample of size n_A taken from frame A is denoted by $\mathcal{S}_A$, and the first-order inclusion probabilities are denoted by $\pi_{Ai} = P(i \in \mathcal{S}_A)$; $\mathcal{S}_B$, π_{Bi}, and n_B are defined similarly for the sample from frame B. The two samples $\mathcal{S}_A$ and $\mathcal{S}_B$ are independent. Frame A sample $\mathcal{S}_A$ can be poststratified as $\mathcal{S}_A = \mathcal{S}_a \cup \mathcal{S}_{ab}$ over the two domains a and ab, where $\mathcal{S}_a = a \cap \mathcal{S}_A$ and $\mathcal{S}_{ab} = \mathcal{S}_A \cap (ab)$. Similarly, frame B sample $\mathcal{S}_B$ can be poststratified as $\mathcal{S}_B = \mathcal{S}_b \cup \mathcal{S}_{ba}$ over the two domains b and ab, where $\mathcal{S}_b = b \cap \mathcal{S}_B$ and $\mathcal{S}_{ba} = \mathcal{S}_B \cap (ab)$. Note that both $\mathcal{S}_{ab}$ and $\mathcal{S}_{ba}$ are from the common domain ab, but $\mathcal{S}_{ab}$ is part of the frame A sample and $\mathcal{S}_{ba}$ is part of the frame B sample. This notation differs from the notation $\mathcal{S}'_{ab}$ and $\mathcal{S}''_{ab}$ used by Hartley (1962) and others.

If no auxiliary information is involved at the estimation stage and if the goal is to obtain a point estimator of $\bar{Y}$, then the problem essentially reduces to estimating the domain mean $\bar{Y}_{ab}$ using two independent samples $\mathcal{S}_{ab}$ and $\mathcal{S}_{ba}$, plus an estimator of $\bar{Y}_a$ using $\mathcal{S}_a$ and an estimator of $\bar{Y}_b$ using $\mathcal{S}_b$. The final estimator of $\bar{Y}$ can then be obtained using (2.1). The PEL approach that we propose here will simultaneously achieve three major goals: (1) obtain a point estimator of $\bar{Y}$, (2) incorporate auxiliary population information, (3) construct confidence intervals on $\bar{Y}$.

Although $\mathcal{S}_{ab}$ and $\mathcal{S}_{ba}$ should be viewed as two independent samples from the same domain ab, it is strategically more convenient to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, and to view $\mathcal{S}_{ab}$ as a sample from ab and $\mathcal{S}_{ba}$ as a sample from ba. We can then rewrite $\bar{Y}$ as an overall population mean over four strata, i.e.,

$$\bar{Y} = W_a\bar{Y}_a + W_{ab}(\eta)\bar{Y}_{ab} + W_{ba}(\eta)\bar{Y}_{ba} + W_b\bar{Y}_b,$$

where $W_a = N_a/N$, $W_{ab}(\eta) = \eta N_{ab}/N$, $W_{ba}(\eta) = (1-\eta)N_{ab}/N$, $W_b = N_b/N$, $\bar{Y}_{ba} = \bar{Y}_{ab}$, and $\eta \in (0,1)$ is a fixed constant to be specified. Note that $W_{ab}(\eta)\bar{Y}_{ab} + W_{ba}(\eta)\bar{Y}_{ba} = (N_{ab}/N)\bar{Y}_{ab}$ for any η. The dual-frame samples $\mathcal{S}_A$ and $\mathcal{S}_B$ can be simply combined into a single "poststratified sample" $(\mathcal{S}_a, \mathcal{S}_{ab}, \mathcal{S}_{ba}, \mathcal{S}_b)$, with "poststratum" sample sizes $(n_a, n_{ab}, n_{ba}, n_b)$. Note that $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$.

Following the stratified formulation of the PEL approach described by Wu and Rao (2006), we define the PEL function for dual-frame samples as

$$\begin{aligned} l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b) &= n_D\Bigg\{W_a \sum_{i\in\mathcal{S}_a} \tilde{d}_{ai}(\mathcal{S}_a)\log(p_{ai}) \\ &\quad + W_{ab}(\eta) \sum_{i\in\mathcal{S}_{ab}} \tilde{d}_{abi}(\mathcal{S}_{ab})\log(p_{abi}) \\ &\quad + W_{ba}(\eta) \sum_{i\in\mathcal{S}_{ba}} \tilde{d}_{bai}(\mathcal{S}_{ba})\log(p_{bai}) \\ &\quad + W_b \sum_{i\in\mathcal{S}_b} \tilde{d}_{bi}(\mathcal{S}_b)\log(p_{bi})\Bigg\}, \end{aligned} \tag{2.2}$$

where $n_D = n_A + n_B$, $\tilde{d}_{ai}(\mathcal{S}_a) = d_{Ai}/\sum_{i\in\mathcal{S}_a} d_{Ai}$, $\tilde{d}_{abi}(\mathcal{S}_{ab}) = d_{Ai}/\sum_{i\in\mathcal{S}_{ab}} d_{Ai}$, $d_{Ai} = 1/\pi_{Ai}$; $\tilde{d}_{bi}(\mathcal{S}_b)$ and $\tilde{d}_{bai}(\mathcal{S}_{ba})$ are defined similarly. The four sets of probability measures, $\mathbf{p}_a =$

$(p_{a1},\ldots,p_{an_a})'$, $\mathbf{p}_{ab} = (p_{ab1},\ldots,p_{abn_{ab}})'$, $\mathbf{p}_{ba} = (p_{ba1},\ldots,p_{ban_{ba}})'$, and $\mathbf{p}_b = (p_{b1},\ldots,p_{bn_b})'$, correspond to poststratified samples $\mathcal{S}_a$, $\mathcal{S}_{ab}$, $\mathcal{S}_{ba}$, and $\mathcal{S}_b$, respectively.

The set of normalization constraints under the current formulation is specified as

$$\begin{aligned} &\sum_{i\in\mathcal{S}_a} p_{ai} = 1, \qquad \sum_{i\in\mathcal{S}_{ab}} p_{abi} = 1, \\ &\sum_{i\in\mathcal{S}_{ba}} p_{bai} = 1, \qquad \text{and} \qquad \sum_{i\in\mathcal{S}_b} p_{bi} = 1. \end{aligned} \tag{2.3}$$

The constraint induced by the common domain mean $\bar{Y}_{ba} = \bar{Y}_{ab}$ is given by

$$\sum_{i\in\mathcal{S}_{ab}} p_{abi}y_i = \sum_{j\in\mathcal{S}_{ba}} p_{baj}y_j. \tag{2.4}$$

The maximum PEL estimator of $\bar{Y}$ based on this poststratified formulation is computed as

$$\hat{\bar{Y}}_P = W_a\hat{\bar{Y}}_a + W_{ab}(\eta)\hat{\bar{Y}}_{ab} + W_{ba}(\eta)\hat{\bar{Y}}_{ba} + W_b\hat{\bar{Y}}_b, \tag{2.5}$$

where $\hat{\bar{Y}}_a = \sum_{i\in\mathcal{S}_a}\hat{p}_{ai}y_i$, $\hat{\bar{Y}}_{ab} = \sum_{i\in\mathcal{S}_{ab}}\hat{p}_{abi}y_i$, $\hat{\bar{Y}}_{ba} = \sum_{i\in\mathcal{S}_{ba}}\hat{p}_{bai}y_i = \hat{\bar{Y}}_{ab}$ due to constraint (2.4), and $\hat{\bar{Y}}_b = \sum_{i\in\mathcal{S}_b}\hat{p}_{bi}y_i$, with $\hat{p}_{ai}, \hat{p}_{abi}, \hat{p}_{bai}$, and $\hat{p}_{bi}$ maximizing $l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ subject to constraints (2.3) and (2.4). Note that $\hat{\bar{Y}}_P$ also can be written as $\hat{\bar{Y}}_P = W_a\hat{\bar{Y}}_a + (N_{ab}/N)\hat{\bar{Y}}_{ab} + W_b\hat{\bar{Y}}_b$, and the choice of η only affects the estimator $\hat{\bar{Y}}_{ab}$ for the population mean of the overlapping domain, $\bar{Y}_{ab}$.

Proposition 1. Under regularity conditions C1–C3 specified in Section A.1 and assuming that $n_{ab} \to \infty$, $n_{ba} \to \infty$ and $n_{ab}/(n_{ab}+n_{ba}) \to c_0 \in (0,1)$ as $n_A \to \infty$ and $n_B \to \infty$, we have $\hat{\bar{Y}}_{ab} = \eta\hat{\bar{Y}}_{abH} + (1-\eta)\hat{\bar{Y}}_{baH} + o_p(m^{-1/2})$, where $\hat{\bar{Y}}_{abH} = \sum_{i\in\mathcal{S}_{ab}}\tilde{d}_{abi}(\mathcal{S}_{ab})y_i$ and $\hat{\bar{Y}}_{baH} = \sum_{i\in\mathcal{S}_{ba}}\tilde{d}_{bai}(\mathcal{S}_{ba})y_i$ are the "Hájek estimators" of $\bar{Y}_{ab}$ and $m = \max\{n_{ab}, n_{ba}\}$.

An optimal choice of η that minimizes the asymptotic variance of $\hat{\bar{Y}}_{ab}$ is given by

$$\eta_o = V_B(\hat{\bar{Y}}_{baH})/\{V_A(\hat{\bar{Y}}_{abH}) + V_B(\hat{\bar{Y}}_{baH})\}, \tag{2.6}$$

where V_A and V_B denote variances under frame A and frame B sampling designs, respectively. A consistent estimator $\hat{\eta}_o$ can be obtained by substituting in (2.6) consistent variance estimators $v_A(\hat{\bar{Y}}_{abH})$ and $v_B(\hat{\bar{Y}}_{baH})$. Under regularity condition C2 specified in Section A.1, we have $\hat{\bar{Y}}_{abH} = \bar{Y}_{ab} + N_{ab}^{-1}\sum_{i\in\mathcal{S}_A} d_{Ai}z_{Ai} + o_p(n_A^{-1/2})$, where $z_{Ai} = y_i - \bar{Y}_{ab}$ if $i \in \mathcal{S}_{ab}$ and $z_{Ai} = 0$ if $i \in \mathcal{S}_a$, and a linearization variance estimator $v_A(\hat{\bar{Y}}_{abH})$ is readily obtained. Similar procedures can be applied to obtain $v_B(\hat{\bar{Y}}_{baH})$.

The asymptotic optimality remains valid when η_o is replaced by $\hat{\eta}_o$, because $\{\hat{\eta}_o\hat{\bar{Y}}_{abH} + (1-\hat{\eta}_o)\hat{\bar{Y}}_{baH}\} = \{\eta_o\hat{\bar{Y}}_{abH} + (1-\eta_o)\hat{\bar{Y}}_{baH}\} + o_p(m^{-1/2})$. Noting that $\hat{\bar{Y}}_a = \hat{\bar{Y}}_{aH}$ and $\hat{\bar{Y}}_b = \hat{\bar{Y}}_{bH}$ are both Hájek estimators, we have

$$\begin{aligned} V(\hat{\bar{Y}}_P) \doteq {}& V_A\{W_a\hat{\bar{Y}}_{aH} + W_{ab}(\eta_o)\hat{\bar{Y}}_{abH}\} \\ & + V_B\{W_b\hat{\bar{Y}}_{bH} + W_{ba}(\eta_o)\hat{\bar{Y}}_{baH}\}. \end{aligned} \tag{2.7}$$

A linearization variance estimator, $v(\hat{\bar{Y}}_P)$, can be derived along the lines of calculating the design effect presented in Section A.3.

The estimator $\hat{\eta}_o$ depends on the variable y except in the case of simple random sampling from both frames. Skinner and Rao (1996) suggested replacing η_o by

$$\eta_P = N_aN_BV_B(\hat{N}_{ba})/\{N_bN_AV_A(\hat{N}_{ab}) + N_aN_BV_B(\hat{N}_{ba})\}, \tag{2.8}$$

where $\hat{N}_{ab} = \sum_{i\in\mathcal{S}_{ab}} d_{Ai}$ and $\hat{N}_{ba} = \sum_{i\in\mathcal{S}_{ba}} d_{Bi}$ (see also Lohr and Rao 2000, eq. 3). Noting that $\hat{N}_{ab} = \sum_{i\in\mathcal{S}_A} d_{Ai}z_{Ai}$, where $z_{Ai} = 1$ if $i \in \mathcal{S}_{ab}$ and $z_{Ai} = 0$ otherwise, an estimator $v_A(\hat{N}_{ab})$ of $V_A(\hat{N}_{ab})$ can be obtained using a standard variance formula. Similarly, an estimator $v_B(\hat{N}_{ba})$ of $V_B(\hat{N}_{ba})$ can be obtained, leading to a consistent estimator $\hat{\eta}_P$ of η_P. A linearization variance estimator of the resulting estimator, $\hat{\bar{Y}}_P$, can be obtained along the lines of (2.7) by substituting $\hat{\eta}_P$ for η_o.

We now consider the PEL ratio confidence intervals on $\bar{Y}$, treating $\eta = \eta_o$ as fixed. Let $\tilde{\mathbf{p}}_a(\theta)$, $\tilde{\mathbf{p}}_{ab}(\theta)$, $\tilde{\mathbf{p}}_{ba}(\theta)$, and $\tilde{\mathbf{p}}_b(\theta)$ be the maximizer of $l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ under the constrains (2.3), (2.4) and the following additional constraint induced by the parameter of interest, $\bar{Y}$:

$$\begin{aligned} W_a\sum_{i\in\mathcal{S}_a} p_{ai}y_i + W_{ab}(\eta_o)\sum_{i\in\mathcal{S}_{ab}} p_{abi}y_i & \\ + W_{ba}(\eta_o)\sum_{i\in\mathcal{S}_{ba}} p_{bai}y_i + W_b\sum_{i\in\mathcal{S}_b} p_{bi}y_i &= \theta \end{aligned} \tag{2.9}$$

for a fixed θ. The PEL ratio statistic is given by

$$\begin{aligned} r_D(\theta) = -2\big\{& l_D(\tilde{\mathbf{p}}_a(\theta), \tilde{\mathbf{p}}_{ab}(\theta), \tilde{\mathbf{p}}_{ba}(\theta), \tilde{\mathbf{p}}_b(\theta)) \\ & - l_D(\hat{\mathbf{p}}_a, \hat{\mathbf{p}}_{ab}, \hat{\mathbf{p}}_{ba}, \hat{\mathbf{p}}_b)\big\}. \end{aligned} \tag{2.10}$$

Let deff_P be the estimated design effect. We then have the following result concerning the asymptotic distribution of $r_D(\theta)$.

Theorem 1. Under regularity conditions C1–C3 specified in Section A.1, the adjusted PEL ratio statistic $r_D^{[a]}(\theta) = r_D(\theta)/\text{deff}_P$ converges in distribution to a χ^2 random variable with 1 degree of freedom when $\theta = \bar{Y}$.

A sketch of proof of Theorem 1 and details on how to calculate deff_P are given in Section A.3. A $(1-\alpha)$-level confidence interval on $\bar{Y}$ can be constructed as $\mathcal{C}_a = \{\theta \mid r_D^{[a]}(\theta) < \chi_1^2(\alpha)\}$, where $\chi_1^2(\alpha)$ is the $(1-\alpha)$th quantile of the χ^2 distribution with 1 degree of freedom.

One major advantage of the PEL approach is that known auxiliary population information can be incorporated into inference through additional constraints. Suppose that $\bar{\mathbf{X}}_A$ is the vector of known frame A population means of auxiliary variables $\mathbf{x}_A$. This frame-specific information can be incorporated through the constraint

$$\frac{N_a}{N_A}\sum_{i\in\mathcal{S}_a} p_{ai}\mathbf{x}_{Ai} + \frac{N_{ab}}{N_A}\sum_{j\in\mathcal{S}_{ab}} p_{abj}\mathbf{x}_{Aj} = \bar{\mathbf{X}}_A. \tag{2.11}$$

If $\bar{\mathbf{X}}_B$ specific to frame B is also available, then a set of constraints similar to (2.11) can be included as well. A key technical argument used here for both asymptotic development and computational procedures is to reformulate all involved constraints using the formulation for stratified sampling. For this purpose, we rewrite (2.11) as

$$W_a \sum_{i\in\mathcal{S}_a} p_{ai}\mathbf{x}_{Ai} + W_{ab}(\eta_o) \sum_{i\in\mathcal{S}_{ab}} p_{abi}\frac{\mathbf{x}_{Ai}}{\eta_o} + W_{ba}(\eta_o) \sum_{i\in\mathcal{S}_{ba}} p_{bai}\cdot\mathbf{0} + W_b \sum_{i\in\mathcal{S}_b} p_{bi}\cdot\mathbf{0} = \frac{\mathbf{X}_A}{N}, \quad (2.12)$$

where $\mathbf{X}_A = N_A\bar{\mathbf{X}}_A$ is the frame A population total. Suppose that, in addition to frame-specific information, the overall population mean $\bar{\mathbf{X}}$ of $\mathbf{x}$ is also known and the $\mathbf{x}_i$ are observed on both samples $\mathcal{S}_A$ and $\mathcal{S}_B$. This information can be used through the constraints

$$W_a \sum_{i\in\mathcal{S}_a} p_{ai}\mathbf{x}_i + W_{ab}(\eta_o) \sum_{i\in\mathcal{S}_{ab}} p_{abi}\mathbf{x}_i + W_{ba}(\eta_o) \sum_{i\in\mathcal{S}_{ba}} p_{bai}\mathbf{x}_i + W_b \sum_{i\in\mathcal{S}_b} p_{bi}\mathbf{x}_i = \bar{\mathbf{X}} \quad (2.13)$$

and

$$\sum_{i\in\mathcal{S}_{ab}} p_{abi}\mathbf{x}_i = \sum_{i\in\mathcal{S}_{ba}} p_{bai}\mathbf{x}_i. \quad (2.14)$$

Note that constraint (2.14) also can be rewritten as

$$W_a \sum_{i\in\mathcal{S}_a} p_{ai}\cdot\mathbf{0} + W_{ab}(\eta_o) \sum_{i\in\mathcal{S}_{ab}} p_{abi}\frac{\mathbf{x}_i}{\eta_o} + W_{ba}(\eta_o) \sum_{i\in\mathcal{S}_{ba}} p_{bai}\frac{-\mathbf{x}_i}{1-\eta_o} + W_b \sum_{i\in\mathcal{S}_b} p_{bi}\cdot\mathbf{0} = \mathbf{0}. \quad (2.15)$$

For point estimation or computation of the unadjusted PEL ratio function $r_D(\theta)$, we simply include (2.12), (2.13), and (2.15) as part of the maximization process. For PEL ratio confidence intervals where the design effect deff_P is needed, we augment the $\mathbf{z}_i$ variable defined in Section A.3 to include variables appearing on the left side of those equations and augment $\bar{\mathbf{Z}}$ to include variables appearing on the right side of those equations. For instance, if constraint (2.12) is included, then $\bar{\mathbf{Z}} = (W_a, W_{ab}(\eta_o), W_{ba}(\eta_o), 0, (\mathbf{X}_A/N)')'$ and $\mathbf{z}_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i}, \mathbf{z}_{5i}')'$, where z_{hi}, $h = 1, 2, 3, 4$, are as defined in Section A.3 and $\mathbf{z}_{5i} = \mathbf{x}_{Ai}$ if $i \in \mathcal{S}_a$, $\mathbf{z}_{5i} = \mathbf{x}_{Ai}/\eta_o$ if $i \in \mathcal{S}_{ab}$, and $\mathbf{z}_{5i} = \mathbf{0}$ if $i \in \mathcal{S}_{ba}$ or $i \in \mathcal{S}_b$.

A practically important application of Case 1 occurs when frame A is complete, frame B is incomplete, and both N_A and N_B are known. In this case $N_{ab} = N_B$, $N_a = N_A - N_B$, and $N_b = 0$. Consequently, $W_b = 0$, $\mathcal{S}_{ba} = \mathcal{S}_B$, and $\mathcal{S}_b = \emptyset$. All terms involving $\mathcal{S}_b$ disappear from previous formulas, including the definition of the PEL function $l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$.

Case 2: N_A and N_B Known but N_{ab} Unknown

The unknown N_{ab} can be estimated by $\hat{N}_{ab,P} = \hat{\phi}\hat{N}_{ab} + (1-\hat{\phi})\hat{N}_{ba}$, where $\hat{N}_{ab} = \sum_{i\in\mathcal{S}_{ab}} d_{Ai}$, $\hat{N}_{ba} = \sum_{i\in\mathcal{S}_{ba}} d_{Bi}$, and $\hat{\phi} = v_B(\hat{N}_{ba})/\{v_A(\hat{N}_{ab}) + v_B(\hat{N}_{ba})\}$ (see Lohr and Rao 2000, eq. 4). For asymptotic development, $\hat{\phi}$ can be replaced by $\phi = V_B(\hat{N}_{ba})/\{V_A(\hat{N}_{ab}) + V_B(\hat{N}_{ba})\}$; that is, $\hat{\phi}$ can be viewed as fixed. Under the current case, the PEL function $l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ as well as constraints involved must be modified from Case 1 by making the following changes: Replace W_a, $W_{ab}(\eta_o)$, $W_{ba}(\eta_o)$, and W_b by $\hat{W}_a = \hat{N}_{a,P}/\hat{N}_P$, $\hat{W}_{ab}(\eta_o) = \eta_o\hat{N}_{ab,P}/\hat{N}_P$, $\hat{W}_{ba}(\eta_o) = (1-\eta_o)\hat{N}_{ab,P}/\hat{N}_P$, and $\hat{W}_b = \hat{N}_{b,P}/\hat{N}_P$, respectively, where $\hat{N}_P = N_A + N_B - \hat{N}_{ab,P}$, $\hat{N}_{a,P} = N_A - \hat{N}_{ab,P}$, and $\hat{N}_{b,P} = N_B - \hat{N}_{ab,P}$.

It can be seen that all technical arguments under the current setting parallel those of Case 1, with the only major difference involving variance estimation and calculation of the design effect, deff_P. The maximum PEL estimator of $\bar{Y}$ is given by $\hat{\bar{Y}}_P = \hat{Y}_P/\hat{N}_P$, where $\hat{Y}_P = (N_A - \hat{N}_{ab,P})\hat{\bar{Y}}_a + \hat{N}_{ab,P}\hat{\bar{Y}}_{ab} + (N_B - \hat{N}_{ab,P})\hat{\bar{Y}}_b$. A tedious but straightforward linearization procedure leads to

$$V(\hat{\bar{Y}}_P) \doteq N^{-2}\left\{V_A\left(\sum_{i\in\mathcal{S}_A} d_{Ai}z_{Ai}\right) + V_B\left(\sum_{i\in\mathcal{S}_B} d_{Bi}z_{Bi}\right)\right\}, \quad (2.16)$$

where $z_{Ai} = y_i - \bar{Y}_a$ if $i \in \mathcal{S}_a$, $z_{Ai} = \eta_o(y_i - \bar{Y}_{ab}) + \phi k$ if $i \in \mathcal{S}_{ab}$, $z_{Bi} = y_i - \bar{Y}_b$ if $i \in \mathcal{S}_b$, $z_{Bi} = (1-\eta_o)(y_i - \bar{Y}_{ab}) + (1-\phi)k$ if $i \in \mathcal{S}_{ba}$, and $k = \bar{Y} - \bar{Y}_a - \bar{Y}_b + \bar{Y}_{ab}$. The final variance estimator, $v(\hat{\bar{Y}}_P)$, then can be derived based on (2.16).

The variance part involved in calculating the design effect deff_P for Case 2 (outlined in Section A.3 for Case 1) is now given by $V(\sum_{h=1}^4 \hat{W}_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i) \doteq N^{-2}\{V_A(\sum_{i\in\mathcal{S}_A} d_{Ai}z_{Ai}) + V_B(\sum_{i\in\mathcal{S}_B} d_{Bi}z_{Bi})\}$, where z_{Ai} and z_{Bi} are defined similarly as in (2.16), with y_i replaced by r_i and $\bar{Y}$, $\bar{Y}_a$, $\bar{Y}_b$, and $\bar{Y}_{ab}$ replaced by the corresponding population means for the r-variable.

Case 3: N_A, N_B, and N_{ab} All Unknown

The unknown frame population sizes N_A and N_B can be estimated by $\hat{N}_A = \sum_{i\in\mathcal{S}_A} d_{Ai}$ and $\hat{N}_B = \sum_{i\in\mathcal{S}_B} d_{Bi}$, N_{ab} is estimated by $\hat{N}_{ab,P}$ as in Case 2, and N is estimated by $\hat{N}_P = \hat{N}_A + \hat{N}_B - \hat{N}_{ab,P}$. The maximum PEL estimator of $\bar{Y}$ is now given by $\hat{\bar{Y}}_P = \hat{Y}_P/\hat{N}_P$, where $\hat{Y}_P = (\hat{N}_A - \hat{N}_{ab,P})\hat{\bar{Y}}_a + \hat{N}_{ab,P}\hat{\bar{Y}}_{ab} + (\hat{N}_B - \hat{N}_{ab,P})\hat{\bar{Y}}_b$. A linearization variance estimator can be derived based on (2.16), where $z_{Ai} = y_i - \bar{Y}$ if $i \in \mathcal{S}_a$, $z_{Ai} = \eta_o(y_i - \bar{Y}_{ab}) + (1-\phi)(\bar{Y}_a - \bar{Y}) + \phi(\bar{Y}_{ab} - \bar{Y}_b)$ if $i \in \mathcal{S}_{ab}$, $z_{Bi} = y_i - \bar{Y}$ if $i \in \mathcal{S}_b$, $z_{Bi} = (1-\eta_o)(y_i - \bar{Y}_{ab}) + \phi(\bar{Y}_b - \bar{Y}) + (1-\phi)(\bar{Y}_{ab} - \bar{Y}_a)$ if $i \in \mathcal{S}_{ba}$. The estimated design effect deff_P can be computed similarly.

3. A SINGLE–FRAME MULTIPLICITY–BASED APPROACH TO MULTIPLE–FRAME SURVEYS

In this section we present a single-frame multiplicity-based PEL approach to inferences from multiple-frame surveys. The method requires less information on frame membership details and is generally applicable to Q-frame ($Q \ge 2$) survey samples.

Suppose that there are Q frames, denoted by $A_1, \ldots, A_Q$, $Q \ge 2$. These frames make a partition of the overall population into possibly $2^Q - 1$ nonoverlapping domains. All existing approaches, including the PEL method presented in Section 2, require that the domain membership be correctly identified for all sampled units. This information is not always available, however. A partial membership information, termed *multiplicity*, which is *the number of frames* to which a particular unit belongs, often can be obtained with some minor effort during the data collection process (Mecatti 2007).

Let $\mathcal{S}_1, \ldots, \mathcal{S}_Q$ be Q independent samples drawn from the Q frames. Let $\{(y_{qi}, \mathbf{x}_{qi}), i \in \mathcal{S}_q\}$, $q = 1, \ldots, Q$ be the Q-frame survey samples, where y_{qi} is the common response variable attached to unit i on frame A_q and $\mathbf{x}_{qi}$ is a vector of auxiliary variables that are not necessarily common across different frames. Let $d_{qi} = 1/\pi_{qi}$ be the design weights associated with frame A_q, where $\pi_{qi} = P(i \in \mathcal{S}_q)$ are the first-order inclusion probabilities for the frame A_q sampling design. Let $\mathcal{A}_q$ be the set of all units in frame A_q and $\mathcal{U} = \{1, 2, \ldots, N\}$ be the complete set of units for the overall finite population of size N. Any frame-specific unit (qi) corresponds to a unique $j \in \mathcal{U}$. The key concept here is the so-called multiplicity, m_{qi}, defined as the number of frames to which unit i in frame A_q belongs. For dual-frame surveys, $m_{qi} = 1$ if $i \in a$ or $i \in b$ and $m_{qi} = 2$ if $i \in ab$. For $Q > 2$, this information is less demanding than the specific domain membership and may be possible to obtain without much difficulty. It is straightforward to show that $\sum_{q=1}^{Q} \sum_{i \in \mathcal{A}_q} y_{qi}/m_{qi} = \sum_{j=1}^{N} y_j = Y$. A design-unbiased estimator of the population total Y is given by

$$\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} d_{qi} \frac{y_{qi}}{m_{qi}}. \tag{3.1}$$

The foregoing approach is equivalent to pooling together the Q frames into a single frame that keeps all duplicated units and replacing y_{qi} by y_{qi}/m_{qi}. This amounts to letting the value of response variable y_{qi} be shared by the m_{qi} frames to which unit qi belongs. The idea of *variable sharing* was first used by Rao (1968) to handle a single frame with an unknown amount of duplication. An unbiased estimator of the overall population size N is given by $\hat{N}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} d_{qi}/m_{qi}$, and the Hájek estimator of the population mean is given by $\hat{\bar{Y}}_H = \hat{Y}_M/\hat{N}_M$. The single-frame *variable sharing* estimator $\hat{Y}_M$, given in (3.1), also can be viewed from a different angle. If we rewrite the estimator as $\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} (d_{qi}/m_{qi}) y_{qi}$, then it is the so-called *weight sharing* estimator (Lavallee 2007). The basic design weight d_{qi} attached to unit i in frame A_q is shared by the same unit on all m_{qi} different frames.

We define the single-frame multiplicity-based PEL function for the Q-frame survey samples as

$$l_M(\mathbf{p}_1, \ldots, \mathbf{p}_Q) = \frac{n_M}{\hat{N}_M} \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} \frac{d_{qi}}{m_{qi}} \log(p_{qi}), \tag{3.2}$$

where $n_M = \sum_{q=1}^{Q} n_q$, and n_q is the size of sample $\mathcal{S}_q$ from frame A_q, $\mathbf{p}_q = (p_{q1}, \ldots, p_{qn_q})'$ is the set of probability measures attached to sample $\mathcal{S}_q$, $q = 1, \ldots, Q$. Ignoring the multiplying constant $n_M/\hat{N}_M$, $l_M(\mathbf{p}_1, \ldots, \mathbf{p}_Q)$ is a design-unbiased estimator of the census log-likelihood, $\sum_{i=1}^{N} \log(p_i)$. In the absence of auxiliary population information, maximizing $l_M(\mathbf{p}_1, \ldots, \mathbf{p}_Q)$ subject to

$$\sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} p_{qi} = 1 \tag{3.3}$$

gives $\hat{p}_{qi} = (d_{qi}/m_{qi})/\hat{N}_M$. The maximum PEL estimator of $\bar{Y}$, computed as $\hat{\bar{Y}}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} \hat{p}_{qi} y_{qi}$, is identical to the Hájek estimator $\hat{\bar{Y}}_H = \hat{Y}_M/\hat{N}_M$. For dual-frame surveys, the estimator $\hat{\bar{Y}}_M$ is asymptotically equivalent to the poststratified maximum PEL estimator $\hat{\bar{Y}}_P$ for the case of N_A, N_B, and N_{ab} all unknown and the choice of $\eta = \phi = 1/2$. Thus the multiplicity-based estimator is not necessarily optimal under dual-frame surveys. The major advantage of the approach is its simplicity in handling general Q-frame surveys.

Let the $\tilde{p}_{qi}(\theta)$ be the maximizers of $l_M(\mathbf{p}_1, \ldots, \mathbf{p}_Q)$ under the constraint (3.3) and

$$\sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} p_{qi} y_{qi} = \theta \tag{3.4}$$

for a fixed θ. The multiplicity-based PEL ratio function for the population mean $\bar{Y}$ is defined as

$$r_M(\theta) = -2\big\{l_M(\tilde{\mathbf{p}}_1(\theta), \ldots, \tilde{\mathbf{p}}_Q(\theta)) - l_M(\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_Q)\big\}. \tag{3.5}$$

Let deff_M be the design effect associated with $\hat{\bar{Y}}_H$; see Section A.4 for details. Let C1* and C2* be regularity conditions similar to C1 and C2 given in Section A.1 but extended from dual frames to multiple frames. We then have the following result concerning the asymptotic distribution of $r_M(\theta)$.

Theorem 2. Under regularity conditions C1* and C2*, the adjusted PEL ratio statistic $r_M^{[a]}(\theta) = r_M(\theta)/\text{deff}_M$ converges in distribution to a χ^2 random variable with 1 degree of freedom when $\theta = \bar{Y}$.

The proof of Theorem 2 is similar to that of theorem 1 of Wu and Rao (2006) for single-frame surveys and is omitted here. Note that Theorem 2 holds even if some of the frames involved are complete, because whether a frame is complete or incomplete has no direct consequences on conditions C1* and C2*. If the vector of the overall population mean $\bar{\mathbf{X}}$ of common auxiliary variables $\mathbf{x}$ is known and $\mathbf{x}_{qi}$ is observed in all samples $\mathcal{S}_q$, $q = 1, \ldots, Q$, then this information can be conveniently used through the constraint

$$\sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} p_{qi} \mathbf{x}_{qi} = \bar{\mathbf{X}}. \tag{3.6}$$

A result similar to Theorem 2 can be established when constraint (3.6) is included in calculating both $\hat{p}_{qi}$ and $\tilde{p}_{qi}(\theta)$. The design effect in this case is denoted as $\text{deff}_{\text{GR}(M)}$ and is related to a generalized regression estimator of $\bar{Y}$.

4. BOOTSTRAP CALIBRATED PEL RATIO CONFIDENCE INTERVALS

Construction of PEL ratio confidence intervals based on Theorems 1 and 2 requires consistent estimation of the design effects deff_P and deff_M, which involves variance estimation, as detailed in Sections A.3 and A.4. This problem can be alleviated through a bootstrap calibration method. Wu and Rao (2010) described bootstrap procedures for the PEL method involving single-frame surveys. Their proposed procedure for stratified sampling is directly applicable to dual-frame surveys under the poststratified formulation for the case of N_A, N_B, and N_{ab} all known. Simulation results seem to indicate that the procedure works for the case of unknown N_{ab} as well.

For the single-frame multiplicity-based PEL approach to multiple-frame surveys, the bootstrap procedure of Wu and Rao (2010) for nonstratified sampling designs can be applied with some minor modifications. First, both the design weights d_{qi} and the multiplicity m_{qi} need to be treated as part of the qth frame sample data. Bootstrap samples selected from the qth frame sample are in the form of $\{(d_{qi}^*, m_{qi}^*, y_{qi}^*, \mathbf{x}_{qi}^*), i \in \mathcal{S}_q^*\}$, where $\mathcal{S}_q^*$ is a set of n_q units selected from $\mathcal{S}_q$ using simple random sampling with replacement. Second, the bootstrap version of the PEL function specified in (3.2) is given by

$$l_M^*(\mathbf{p}_1, \dots, \mathbf{p}_Q) = \frac{n_M}{\hat{N}_M^*} \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q^*} \frac{d_{qi}^*}{m_{qi}^*} \log(p_{qi}), \tag{4.1}$$

where $\hat{N}_M^* = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q^*} d_{qi}^* / m_{qi}^*$. Bootstrap versions of the constraints (3.4) and (3.6) can be defined similarly as

$$\sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q^*} p_{qi} y_{qi}^* = \hat{\bar{Y}}_M \tag{4.2}$$

and

$$\sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q^*} p_{qi} \mathbf{x}_{qi}^* = \bar{\mathbf{X}}. \tag{4.3}$$

A bootstrap calibrated PEL ratio confidence interval on $\bar{Y}$ is constructed as $\mathcal{C}_u^* = \{\theta | r_M(\theta) < b_\alpha^*\}$, where b_α^* is the $(1-\alpha)$th quantile of the unadjusted PEL ratio statistic $r_M(\theta)$ at $\theta = \bar{Y}$ and can be approximated by the $(1-\alpha)$th sample quantile of $(r_M^{[1]}(\bar{Y}), \dots, r_M^{[K]}(\bar{Y}))$ obtained from K independent bootstrap samples.

The foregoing bootstrap calibration method bypasses the need for calculating the design effects and hence avoids variance estimation. The method is valid for single-stage unequal-probability sampling designs for all frames with small sampling fractions. When a multistage clustering sampling design is used for any of the sampling frames, bootstrap procedures are not readily available, and further research is needed.

5. SIMULATION STUDIES

We conducted an extensive simulation study to examine the finite-sample performances of the proposed PEL methods for two- and three-frame surveys. Both point estimators and confidence intervals were considered. In particular, we considered single-stage unequal probability sampling designs and included the bootstrap-calibrated PEL ratio confidence intervals as part of the study. Our simulations were programmed in R using algorithms developed by Chen, Sitter, and Wu (2002) and Wu (2004, 2005).

We first considered dual-frame surveys using a synthetic finite population created from Statistics Canada's 2000 Family Expenditure Survey in the province of Ontario. The original data set contains 2396 sampled households with measures on y: total expenditure, x_1: total income, x_2: number of people in the household, x_3: number of children (under 15 years old), and several other variables. We created frame A population by including all households with at least one child. This frame may be viewed as the list of households on government's child tax benefit program. Frame B population consisted of households with no more than three people. The resulting population sizes for the two frames are $N_A = 1007$ and $N_B = 1724$, and the overall population size is $N = 2248$. The three domain sizes are $N_a = 524$, $N_b = 1241$, and $N_{ab} = 483$.

Samples from frame A were selected by the Rao–Sampford PPS sampling method (Rao 1965; Sampford 1967), with inclusion probabilities proportional to the household's total income (x_1). Samples from frame B were taken by simple random sampling without replacement. Our focus is on estimating the population mean, $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$, of household's total expenditure, y. The population correlation coefficient between the response variable y, and the design variable x_1 is 0.8. The PPS sampling design used here is highly efficient in estimating $\bar{Y}$, and the design effect over simple random sampling is larger than 3. This is a common scenario for dual-frame surveys where the two sampling designs are often very different.

Among existing estimators for dual-frame surveys, the optimal pseudo–maximum likelihood (PML) estimator $\hat{\bar{Y}}_{PML}$, proposed by Skinner and Rao (1996), is the most competitive one. It is applicable to surveys with complex designs. We included this estimator in the simulation for the purpose of comparison.

For each simulated dual-frame sample $(\mathcal{S}_A, \mathcal{S}_B)$ of size (n_A, n_B), we computed five point estimators of $\bar{Y}$: (1) PML(N_{ab}): the optimal PML estimator, assuming the domain population size N_{ab} is known, (2) PML($\hat{N}_{ab}$): the optimal PML estimator, using the estimated domain size $\hat{N}_{ab}$ given by Skinner and Rao (1996), (3) PEL(N_{ab}): the maximum PEL estimator based on the poststratified formulation, assuming N_{ab} is known, (4) PEL($\hat{N}_{ab}$): the maximum PEL estimator with N_{ab} estimated by $\hat{N}_{ab,P}$, (5) PEL(M): the maximum PEL estimator under the multiplicity-based formulation. For the two poststratified PEL estimators PEL(N_{ab}) and PEL($\hat{N}_{ab}$), the asymptotically optimal estimator $\hat{\eta}_o$ is used. Note that N_{ab} is not required for PEL(M).

Frame A sample sizes were chosen as $n_A = 50$ and 100, corresponding to frame A sampling fractions of 5% and 10%. Frame B sample sizes were chosen at three levels as $n_B = 50$, 100, and 150, corresponding to sampling fractions 3%, 6%, and 9%. Simulated relative bias (RB%) and mean squared error (MSE) for the five estimators and five combinations of the sample sizes (n_A, n_B) are presented in Table 1, based on 1000 simulation runs. Table 1 shows that all five estimators are virtually unbiased, with the largest absolute relative bias equal to 0.7%. For a given sample size (n_A, n_B), MSE's of the five estimators are also very close to one another.

We now turn to the construction of confidence intervals on $\bar{Y}$. For each simulated dual-frame sample of size (n_A, n_B), we computed five confidence intervals: (1) PML($\hat{N}_{ab}$), NA: the interval based on normal approximations (NA) using the PML estimator and its estimated variance (Skinner and Rao 1996, sec. 6), (2) PEL($\hat{N}_{ab}$), χ^2: the PEL ratio confidence interval based on the asymptotic χ^2 distribution of the adjusted PEL ratio function under the poststratified formulation (Theorem 1 of Section 2), (3) PEL($\hat{N}_{ab}$), BT: the PEL ratio confidence interval based on bootstrap (BT) calibration to the unadjusted PEL ratio function under the poststratified formulation, (4) PEL(M), χ^2: the multiplicity-based PEL ratio confidence interval using the asymptotic χ^2 distribution of the adjusted PEL ratio function (Theorem 2 of Section 3), (5) PEL(M), BT: the multiplicity-based PEL ratio confidence interval using bootstrap calibration to the unadjusted PEL ratio function. The estimated domain

Table 1. Simulated RB% and MSE of dual-frame estimators

n_A	n_B		PML(N_{ab})	PML($\hat{N}_{ab}$)	PEL(N_{ab})	PEL($\hat{N}_{ab}$)	PEL(M)
50	50	RB%	−0.7	−0.4	−0.6	−0.4	0.0
		MSE	10.1	9.9	10.6	10.3	10.1
100	50	RB%	−0.6	−0.4	−0.4	−0.3	−0.4
		MSE	9.0	8.9	9.2	9.0	9.3
50	100	RB%	−0.5	−0.3	−0.7	−0.5	−0.1
		MSE	6.6	6.6	7.0	6.8	6.7
100	100	RB%	−0.4	−0.3	−0.4	−0.3	−0.1
		MSE	4.9	4.8	5.0	4.9	4.6
100	150	RB%	−0.3	−0.2	−0.4	−0.3	0.0
		MSE	3.8	3.7	3.9	3.8	3.4

population size $\hat{N}_{ab}$ is used for the first three intervals and $\hat{\eta}_o$ is used for the second and third intervals.

Table 2 reports the simulated coverage probability (CP), left (L) and right (R) tail error rates, and average length (AL) of the 95% confidence intervals on $\bar{Y}$ based on 1000 simulation runs. Standard errors (SE) for the estimated average length are also included (in parentheses). For the bootstrap calibration method, 1000 bootstrap samples were used for each simulated dual-frame sample. Table 2 shows that all five intervals have coverage probabilities close to the nominal value, but the multiplicity-based PEL ratio confidence intervals are generally shorter and in some cases are much shorter. In addition, the bootstrap calibration method works very well for all scenarios, including cases where one or both sampling fractions are around 10%.

Table 2. Simulated 95% confidence intervals on $\bar{Y}$ for dual-frame surveys

			PML($\hat{N}_{ab}$)	PEL($\hat{N}_{ab}$)	PEL($\hat{N}_{ab}$)	PEL(M)	PEL(M)
n_A	n_B		(NA)	(χ^2)	(BT)	(χ^2)	(BT)
50	50	L	1.5	3.3	2.0	1.9	2.0
		CP	95.9	94.8	95.7	94.4	95.2
		U	2.6	1.9	2.3	3.7	2.8
		AL	13.1	13.1	13.2	12.0	12.3
		(SE)	(0.09)	(0.11)	(0.13)	(0.13)	(0.11)
100	50	L	1.1	2.5	1.7	2.7	2.2
		CP	96.5	95.1	94.8	93.7	94.9
		U	2.4	2.4	3.5	3.6	2.9
		AL	12.2	11.7	11.6	11.5	11.5
		(SE)	(0.09)	(0.11)	(0.12)	(0.10)	(0.10)
50	100	L	1.8	4.6	2.8	2.4	2.4
		CP	96.4	93.5	95.4	95.2	94.8
		U	1.8	1.9	1.8	2.4	2.8
		AL	10.6	10.8	10.7	9.4	9.8
		(SE)	(0.05)	(0.08)	(0.07)	(0.11)	(0.09)
100	100	L	1.0	2.8	2.2	2.6	1.4
		CP	97.6	95.4	95.4	95.0	95.9
		U	1.4	1.8	2.4	2.4	2.7
		AL	9.5	9.3	9.1	8.5	8.7
		(SE)	(0.05)	(0.06)	(0.06)	(0.05)	(0.07)
100	150	L	2.0	3.3	3.5	3.4	2.5
		CP	95.9	95.1	94.5	93.4	94.6
		U	2.1	1.6	2.0	3.2	2.9
		AL	8.2	8.2	8.1	7.2	7.6
		(SE)	(0.03)	(0.05)	(0.05)	(0.03)	(0.07)

The two PEL ratio confidence intervals PEL($\hat{N}_{ab}$) and PEL(M) for the finite population distribution function $F_N(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$, where $I(\cdot)$ is the indicator function perform quite differently under dual-frame sampling designs, however. The multiplicity-based PEL interval performs uniformly better than the interval based on normal approximation, especially when t is in the tail region of the population quantiles. Simulation results (not included here to save space) are similar to those reported in Table 4 under three-frame designs. The PEL interval based on the poststratified sample, on the other hand, is not computable when either $\{I(y_i \le t), i \in \mathcal{S}_{ab}\}$ or $\{I(y_i \le t), i \in \mathcal{S}_{ba}\}$ are all 0's or all 1's. There is a nontrivial probability that this could happen when t is in the tail region and sample sizes are not large.

We now report simulation results based on a three-frame sampling design for the same synthetic population. In addition to the two frames A and B, a third frame C was taken as the complete list of all $N = 2248$ households in the population. Stratified simple random sampling was used for frame C, where the population was stratified into low ($x_1 \le 30{,}000$), middle ($30{,}000 < x_1 < 60{,}000$) and high ($x_1 \ge 60{,}000$) income households. To limit the number of combinations of the three frame sample sizes, equal sample size allocation was used for frame C stratified sampling design. The sampling designs for frames A and B remained the same as in the previous study.

In the first part of the simulation study on three frame designs, we examined two multiplicity-based PEL estimators of the population mean $\bar{Y}$: the Hájek estimator $\hat{\bar{Y}}_H = \hat{Y}_M/\hat{N}_M$, where $\hat{Y}_M$ is given by (3.1) without using any auxiliary population information and the estimator under the constraint (3.6) over household's total income x_1, assuming the population mean $\bar{X}_1$ is known. Simulated relative biases of the two estimators, not included here, are less than 0.1% for all cases considered.

For each simulated three-frame sample of size (n_A, n_B, n_C), we computed five 95% confidence intervals on $\bar{Y}$: (1) M_1(NA): the normal approximation interval based on $\hat{\bar{Y}}_H$ and its estimated variance; (2) $M_1(\chi^2)$: the PEL ratio interval based on Theorem 2 without using any auxiliary population information; (3) M_1(BT): the interval similar to $M_1(\chi^2)$ but using the bootstrap calibration method; (4) $M_2(\chi^2)$: the PEL ratio interval

Table 3. Simulated 95% confidence intervals on $\bar{Y}$ for three-frame surveys

(n_A, n_B, n_C)		M_1(NA)	$M_1(\chi^2)$	M_1(BT)	$M_2(\chi^2)$	M_2(BT)
(50, 50, 60)	L	1.1	1.6	1.3	1.5	1.4
	CP	95.1	95.2	95.2	94.7	95.1
	U	3.8	3.2	3.5	3.8	3.5
	AL	7.8	7.9	7.9	4.8	5.1
(50, 100, 60)	L	2.7	3.2	2.8	2.1	1.7
	CP	95.0	95.1	95.2	94.6	95.5
	U	2.3	1.7	2.0	3.3	2.8
	AL	6.6	6.7	6.7	4.4	4.6
(100, 100, 60)	L	2.3	2.5	2.4	1.9	1.5
	CP	95.0	94.9	95.1	95.3	95.8
	U	2.7	2.6	2.5	2.8	2.7
	AL	6.2	6.3	6.4	4.2	4.4
(50, 50, 120)	L	1.7	2.9	2.7	1.3	1.2
	CP	94.1	93.7	93.8	95.7	95.7
	U	4.2	3.4	3.5	3.0	3.1
	AL	7.2	7.3	7.3	4.3	4.6
(50, 100, 120)	L	2.3	2.6	2.0	2.1	1.8
	CP	94.7	95.1	95.9	93.9	94.8
	U	3.0	2.3	2.1	4.0	3.4
	AL	5.8	5.8	6.0	3.6	3.8
(100, 100, 120)	L	1.5	2.0	1.8	1.9	1.6
	CP	94.8	95.1	95.4	94.9	95.6
	U	3.7	2.9	2.8	3.2	2.8
	AL	5.4	5.5	5.6	3.5	3.6

Table 4. 95% confidence intervals on $F_N(t)$ at $t = t_\alpha$ for three-frame surveys

		$t_{0.05}$	$t_{0.10}$	$t_{0.25}$	$t_{0.50}$	$t_{0.75}$	$t_{0.90}$	$t_{0.95}$
NA	L	0.9	1.5	2.5	3.1	3.8	4.6	7.2
	CP	92.8	92.9	94.6	93.7	94.7	93.9	92.4
	U	6.3	5.6	2.9	3.2	1.5	1.5	0.4
	AL	0.076	0.101	0.129	0.142	0.114	0.073	0.053
	LB	0.014	0.051	0.185	0.430	0.692	0.864	0.924
	UB	0.089	0.152	0.314	0.572	0.807	0.937	0.977
PEL(χ^2)	L	2.8	2.8	2.8	3.1	3.0	2.8	3.8
	CP	95.4	94.3	95.1	93.8	95.4	95.1	93.8
	U	1.8	2.9	2.1	3.1	1.6	2.1	2.4
	AL	0.076	0.101	0.129	0.142	0.114	0.072	0.053
	LB	0.023	0.059	0.189	0.431	0.690	0.860	0.919
	UB	0.099	0.160	0.318	0.572	0.804	0.933	0.972
PEL(BT)	L	3.1	2.7	3.1	3.3	3.3	2.8	3.7
	CP	91.3	94.2	94.6	93.7	95.2	95.0	94.0
	U	5.6	3.1	2.3	3.0	1.5	2.2	2.3
	AL	0.075	0.101	0.128	0.141	0.113	0.072	0.053
	LB	0.023	0.058	0.189	0.431	0.690	0.860	0.919
	UB	0.098	0.159	0.318	0.572	0.803	0.933	0.972

under the additional constraint (3.6) over x_1; (5) M_2(BT): the bootstrap version of $M_2(\chi^2)$. Results based on 1000 simulation runs are reported in Table 3. All five intervals have coverage probabilities close to the nominal value, and the PEL intervals using auxiliary population information are significantly shorter than the intervals without the additional constraint. Bootstrap calibrated PEL ratio confidence intervals have average length and coverage probability comparable to the intervals based on χ^2 approximations but have the advantage of not involving variance estimation under single-stage unequal probability sampling designs.

In the second part of the simulation study on the three frames design, we examined confidence intervals on the population distribution function $F_N(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$. Note that proportions of population units with certain characteristics of interest are special cases of the distribution function. Calculation of point estimators and confidence intervals on $F_N(t)$ at fixed t amounts to replacing y_i by $I(y_i \le t)$ and then following the methods for the population mean. We compared the performances of three confidence intervals on $F_N(t)$: (1) interval based on normal approximation (NA) to the Hájek estimator; (2) PEL ratio interval based on the χ^2 approximation [PEL(χ^2)]; (3) PEL ratio interval using the bootstrap method [PEL(BT)]. No auxiliary population information is involved. Table 4 reports simulated results on the three confidence intervals on $F_N(t)$ with the value of t fixed at five population quantiles. In addition to coverage probabilities (CP), lower (L) and upper (U) tail error rates, and average length (AL), Table 4 also includes average lower bound (LB) and average upper bound (UB) of the intervals. Sample sizes are taken as $n_A = 50$, $n_B = 50$, and $n_C = 60$, with the same sampling designs used for Table 3.

There are two striking observations from Table 4. First, when the values of t are in the middle range of the population quantiles (i.e., $t = t_{0.25}$, $t_{0.50}$ and $t_{0.75}$), all three confidence intervals perform similarly and perform well. This is similar to the observation from Table 3 for the population mean. Second, when t is in the tail region of the population quantiles (i.e., $t = t_{0.05}$ and $t_{0.95}$), PEL ratio intervals clearly outperform normal approximation intervals in terms of coverage probabilities and balanced tail error rates with almost identical average length. For instance, the interval PEL(χ^2) for $F_N(t)$ at $t = t_{0.05}$ has coverage probability of 95.4% compared with 92.8% from the NA interval. The upper and lower tail error rates for the PEL(χ^2) interval are 1.8% and 2.8%, compared with 6.3% and 0.9% from the NA interval. PEL intervals also have larger average lower confidence bounds on $F_N(t)$ when t is a small population quantile and smaller average upper confidence bound when t is a large population quantile. PEL intervals and NA intervals have virtually identical average length. The bootstrap calibrated PEL intervals perform well except for $t = t_{0.05}$, where the coverage probability is low.

6. CONCLUDING REMARKS

Multiple-frame surveys pose several challenges for statistical analysis. Obtaining accurate information on domain membership is the first challenge in practice. Estimating the unknown domain population sizes under complex sampling designs is another problem. Incorporating various auxiliary population information into inferential procedures is also difficult. Variance estimation and confidence intervals are even harder to handle with multiple-frame surveys under general unequal probability sampling designs.

The proposed PEL approach to multiple-frame surveys under a poststratified formulation follows the traditional route in this area. This approach is easy to implement for dual-frame surveys but difficult to extend to three or more frames due to the requirement on domain membership as well as notational complexities. Our proposed multiplicity-based PEL approach, on the other hand, is extremely promising. It is very easy to implement for surveys involving three or more frames, does not require full frame membership information, and yet is flexible in using available auxiliary population information. Multiple-frame surveys are often used to obtain more reliable estimates for population total counts or proportions of rare items, such as people with certain rare disease or illegal status. In this scenario, our proposed PEL ratio confidence intervals have a clear advantage over the customary normal approximation–based intervals, as demonstrated in the simulation study on estimating the population distribution function $F_N(t)$ with t in the tail region of the population quantiles. The PEL approach also has the potential to deal with other types of inferential problems, such as testing of statistical hypothesis or regression analysis using survey data. The required information on multiplicity m_{qi} is also insensitive to domain misclassifications, as shown in the simulation results reported by Mecatti (2007).

The bootstrap-calibrated PEL ratio confidence intervals given in Section 4 are applicable for single-stage unequal probability sampling designs with small sampling fractions. Antal and Tillé (2009) have recently proposed new bootstrap methods for single-frame surveys that work well when sampling fractions are not small. We plan to study PEL intervals for the case of large sampling fractions by adapting their methods to multiple-frame surveys.

The PEL approach to inference for complex surveys is drawing increased attention from survey researchers. An overview of the major theoretical developments as well as computational algorithms for single-frame surveys has been provided by Rao and Wu (2009). The PEL methods presented in this paper for multiple-frame surveys will add a new dimension to the existing literature on the subject.

APPENDIX

A.1 Regularity Conditions for Dual-Frame Surveys

We assume that there is a sequence of dual-frame finite populations, indexed by ν, such that the two frame population sizes, $N_A(\nu)$ and $N_B(\nu)$, and the two sample sizes, $n_A(\nu)$ and $n_B(\nu)$, all tend to infinity as $\nu \to \infty$. If a frame is incomplete, then it remains incomplete as $\nu \to \infty$. The index ν is suppressed for notational simplicity. All limiting processes are understood as $\nu \to \infty$.

C1. Sampling designs for frame A and frame B and the study variable y satisfy $\max_{i\in\mathcal{S}_A} y_i = o_p(n_A^{1/2})$ and $\max_{i\in\mathcal{S}_B} y_i = o_p(n_B^{1/2})$, where the stochastic order $o_p(\cdot)$ is with respect to the sampling design.

C2. The two Horvitz–Thompson estimators $\hat{\theta}_A(z) = N_A^{-1} \times \sum_{i\in\mathcal{S}_A} d_{Ai}z_i$ and $\hat{\theta}_B(z) = N_B^{-1}\sum_{i\in\mathcal{S}_B} d_{Bi}z_i$ are asymptotically normally distributed for any variable z_i such that $\max_{i\in\mathcal{S}_A} z_i = o_p(n_A^{1/2})$ and $\max_{i\in\mathcal{S}_B} z_i = o_p(n_B^{1/2})$.

C3. Suppose that $N_a > 0$ and $N_b > 0$. The asymptotic framework satisfies the conditions $n_A/(n_A + n_B) \to c_1 \in (0, 1)$, $N_a/N_A \to c_2 \in (0, 1)$, and $N_b/N_B \to c_3 \in (0, 1)$ as $n_A \to \infty$ and $n_B \to \infty$.

Condition C1 holds for any sampling designs if $n_A/N_A \to f_A \neq 0$, $n_B/N_B \to f_B \neq 0$ and the finite population values $\{y_1, \dots, y_N\}$ is a random sample from a superpopulation with finite variance. Condition C2 is the central limit theorem for a Horvitz–Thompson estimator (see Wu and Rao 2006, p. 364 for further discussion). Condition C3 does not apply to cases where one of the frames (say frame A) is complete and the other frame (B) is incomplete. In those cases, $N_b = 0$, and the condition $N_b/N_B \to c_3 \in (0, 1)$ must be dropped.

A.2 Proof of Proposition 1

The estimator $\hat{\hat{Y}}_{ab} = \sum_{i\in\mathcal{S}_{ab}} \hat{p}_{abi}y_i = \sum_{i\in\mathcal{S}_{ba}} \hat{p}_{bai}y_i = \hat{\hat{Y}}_{ba}$ is obtained by maximizing the PEL function (2.2) subject to constraints (2.3) and (2.4). It also can be obtained by first replacing (2.4) by $\sum_{i\in\mathcal{S}_{ab}} p_{abi}y_i = \sum_{i\in\mathcal{S}_{ba}} p_{bai}y_i = \theta$ for a fixed θ and then maximizing the resulting PEL function with respect to θ. Suppose that $\hat{\hat{Y}}_{abH} \le \hat{\hat{Y}}_{baH}$ for the given sample. Let $\hat{p}_{ai}(\theta)$, $\hat{p}_{abi}(\theta)$, $\hat{p}_{bai}(\theta)$, and $\hat{p}_{bi}(\theta)$ be the maximizers of $l_D(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ under constraints (2.3) and $\sum_{i\in\mathcal{S}_{ab}} p_{abi}y_i = \sum_{i\in\mathcal{S}_{ba}} p_{bai}y_i = \theta$. If we let $l_D(\theta) = l_D(\hat{\mathbf{p}}_a(\theta), \hat{\mathbf{p}}_{ab}(\theta), \hat{\mathbf{p}}_{ba}(\theta), \hat{\mathbf{p}}_b(\theta))$, then $\hat{\theta} = \hat{\hat{Y}}_{ab}$ is the maximizer of $l_D(\theta)$. To complete the proof, we first show that $\hat{\hat{Y}}_{abH} \le \hat{\theta} \le \hat{\hat{Y}}_{baH}$, which implies that $\hat{\theta} = \bar{Y}_{ab} + O_p(m^{-1/2})$. The desired result on $\hat{\theta}$ follows from solving $\partial l_D(\theta)/\partial\theta = 0$ under standard expansions involving the Lagrange multiplier. Details are omitted.

A.3 Proof of Theorem 1

For notational brevity, we denote (a) W_a, $W_{ab}(\eta_o)$, $W_{ba}(\eta_o)$, and W_b by W_h for $h = 1, 2, 3, 4$; (b) $\mathcal{S}_a$, $\mathcal{S}_{ab}$, $\mathcal{S}_{ba}$, and $\mathcal{S}_b$ by $\mathcal{S}_h$ for $h = 1, 2, 3, 4$; and (c) $\tilde{d}_{ai}(\mathcal{S}_a)$, $\tilde{d}_{abi}(\mathcal{S}_{ab})$, $\tilde{d}_{bai}(\mathcal{S}_{ba})$, and $\tilde{d}_{bi}(\mathcal{S}_b)$ by $\tilde{d}_{hi}$ for $h = 1, 2, 3, 4$. Let $(z_{1i}, z_{2i}, z_{3i})'$ be the vector of indicator variables for $\mathcal{S}_a$, $\mathcal{S}_{ab}$, and $\mathcal{S}_{ba}$. For instance, $z_{1i} = 1$ if $i \in \mathcal{S}_a$ and $z_{1i} = 0$ otherwise. In addition, if we denote p_{ai}, p_{abi}, p_{bai}, and p_{bi} as p_{hi} for $h = 1, 2, 3, 4$, we can rewrite constraint (2.4) as $\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} p_{hi}z_{4i} = 0$, where $z_{4i} = 0$ if $i \in \mathcal{S}_1$ or $i \in \mathcal{S}_4$, $z_{4i} = y_i/\eta_o$ if $i \in \mathcal{S}_2$, and $z_{4i} = -y_i/(1 - \eta_o)$ if $i \in \mathcal{S}_3$. Let $\mathbf{z}_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i})'$ and $\bar{\mathbf{Z}} = (W_a, W_{ab}(\eta_o), W_{ba}(\eta_o), 0)'$. Following along the lines of the proof of theorem 4 of Wu and Rao (2006) and letting $\theta = \bar{Y}$, it can be shown that

$$r_D(\theta) = n_D\left(\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i - \bar{R}\right)^2 \Bigg/ \left\{\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}(r_i - \bar{R})^2\right\} + o_p(1),$$

where $r_i = y_i - \mathbf{B}'\mathbf{z}_i$, $\bar{R} = \bar{Y} - \mathbf{B}'\bar{\mathbf{Z}}$, and $\mathbf{B}$ is the vector of "population regression coefficients" estimated by

$$\hat{\mathbf{B}} = \left\{\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}(\mathbf{z}_i - \bar{\mathbf{Z}})(\mathbf{z}_i - \bar{\mathbf{Z}})'\right\}^{-1} \times \sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}(\mathbf{z}_i - \bar{\mathbf{Z}})(y_i - \hat{\hat{Y}}_P).$$

Under condition C2, $\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i$ is asymptotically normally distributed with mean $\bar{R}$, Theorem 1 is then proved if we define

$$\mathrm{deff}_P = v\left(\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i\right) \Bigg/ \left\{n_D^{-1}\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}(r_i - \bar{r})^2\right\},$$

where $\bar{r} = \hat{\hat{Y}}_P - \hat{\mathbf{B}}'\bar{\mathbf{Z}}$, $v(\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i)$ is an estimator of

$$V\left(\sum_{h=1}^4 W_h \sum_{i\in\mathcal{S}_h} \tilde{d}_{hi}r_i\right) = V_A\{A(r)\} + V_B\{B(r)\},$$

with

$$A(r) = W_a \sum_{i \in \mathcal{S}_a} \tilde{d}_{ai}(\mathcal{S}_a) r_i + W_{ab}(\eta_o) \sum_{i \in \mathcal{S}_{ab}} \tilde{d}_{abi}(\mathcal{S}_{ab}) r_i$$

and

$$B(r) = W_{ba}(\eta_o) \sum_{i \in \mathcal{S}_{ba}} \tilde{d}_{bai}(\mathcal{S}_{ab}) r_i + W_b \sum_{i \in \mathcal{S}_b} \tilde{d}_{bi}(\mathcal{S}_b) r_i.$$

Under condition C2, a standard linearization procedure can be applied to the Hájek estimators $\sum_{i \in \mathcal{S}_a} \tilde{d}_{ai}(\mathcal{S}_a) r_i$ and $\sum_{i \in \mathcal{S}_{ab}} \tilde{d}_{abi}(\mathcal{S}_{ab}) r_i$, which leads to

$$A(r) = W_a \bar{R}_a + W_{ab}(\eta_o) \bar{R}_{ab} + \frac{1}{N} \sum_{i \in \mathcal{S}_A} d_{Ai} \tilde{r}_i + o_p\big(n_A^{-1/2}\big),$$

where $\bar{R}_a$ and $\bar{R}_{ab}$ are the domain population means for the variable r, $\tilde{r}_i = r_i - \bar{R}_a$ if $i \in \mathcal{S}_a$ and $\tilde{r}_i = \eta_o(r_i - \bar{R}_{ab})$ if $i \in \mathcal{S}_{ab}$. This gives $V_A\{A(r)\} \doteq N^{-2} V_A(\sum_{i \in \mathcal{S}_A} d_{Ai} \tilde{r}_i)$. A linearization variance estimator $v_A\{A(r)\}$ can then be derived. Similarly, $V_B\{B(r)\} \doteq N^{-2} V_B(\sum_{i \in \mathcal{S}_B} d_{Bi} \tilde{r}_i)$, where $\tilde{r}_i = r_i - \bar{R}_b$ if $i \in \mathcal{S}_b$ and $\tilde{r}_i = (1 - \eta_o)(r_i - \bar{R}_{ba})$ if $i \in \mathcal{S}_{ba}$.

A.4 Design Effect for the Single-Frame Multiplicity-Based Approach

Here the calculation of design effect involves variance estimation for $\hat{\hat{Y}}_H = \hat{Y}_M / \hat{N}_M$, where $\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} (d_{qi}/m_{qi}) y_{qi}$ and $\hat{N}_M = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} d_{qi}/m_{qi}$. The design effect without involving any auxiliary variable is given by $\text{deff}_M = V(\hat{\hat{Y}}_H)/(S_y^2/n_M)$, where S_y^2 is the overall population variance for the y variable. An approximately unbiased estimator for S_y^2 is given by

$$\hat{S}_y^2 = \frac{1}{\hat{N}_M} \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} \frac{d_{qi}}{m_{qi}} (y_{qi} - \hat{\hat{Y}}_H)^2.$$

The variance estimator for $\hat{\hat{Y}}_H$ is derived as follows:

$$V(\hat{\hat{Y}}_H) \doteq V\left\{ \frac{1}{N} \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}_q} \frac{d_{qi}}{m_{qi}} (y_{qi} - \bar{Y}) \right\} = \frac{1}{N^2} \sum_{q=1}^{Q} V\bigg(\sum_{i \in \mathcal{S}_q} d_{qi} \tilde{y}_{qi} \bigg),$$

where $\tilde{y}_{qi} = (y_{qi} - \bar{Y})/m_{qi}$. Standard variance estimators for the Horvitz–Thompson estimator can be applied to the shared variable $\tilde{y}_{qi}$, with $\bar{Y}$ replaced by $\hat{\hat{Y}}_H$ and N replaced by $\hat{N}_M$ at the final step.

[Received September 2009. Revised April 2010.]

REFERENCES

Antal, E., and Tillé, Y. (2009), “Bootstrap Methods for Complex Sampling Designs in Finite Population,” paper presented at Colloque Deville, Neuchâtel, Switzerland. [1502]

Bankier, M. D. (1986), “Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys,” *Journal of the American Statistical Association*, 81, 1074–1079. [1494]

Chen, J., and Sitter, R. R. (1999), “A Pseudo Empirical Likelihood Approach to the Effective Use of Auxiliary Information in Complex Surveys,” *Statistica Sinica*, 9, 385–406. [1494]

Chen, J., Sitter, R. R., and Wu, C. (2002), “Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys,” *Biometrika*, 89, 230–237. [1499]

Fuller, W. A., and Burmeister, L. F. (1972), “Estimators for Samples Selected From Two Overlapping Frames,” in *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 245–249. [1494]

Hartley, H. O. (1962), “Multiple Frame Surveys,” in *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 203–206. [1494, 1495]

——— (1974), “Multiple Frame Methodology and Selected Applications,” *Sankhyā, Ser. C*, 36, 99–118. [1494,1495]

Kalton, G., and Anderson, D. W. (1986), “Sampling Rare Populations,” *Journal of the Royal Statistical Society, Ser. A*, 149, 65–82. [1494]

Lavallee, P. (2007), *Indirect Sampling*, New York: Springer. [1498]

Lohr, S. L., and Rao, J. N. K. (2000), “Inference From Dual Frame Surveys,” *Journal of the American Statistical Association*, 95, 271–280. [1494-1497]

——— (2006), “Estimation in Multiple-Frame Surveys,” *Journal of the American Statistical Association*, 101, 1019–1030. [1494,1495]

Mecatti, F. (2007), “A Single Frame Multiplicity Estimator for Multiple Frame Survey,” *Survey Methodology*, 33, 151–157. [1497,1502]

Owen, A. B. (1988), “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika*, 75, 237–249. [1494]

Rao, J. N. K. (1965), “On Two Simple Schemes of Unequal Probability Sampling Without Replacement,” *Journal of the Indian Statistical Association*, 3, 173–180. [1499]

——— (1968), “Some Nonresponse Sampling Theory When the Frame Contains an Unknown Amount of Duplication,” *Journal of the American Statistical Association*, 63, 87–90. [1498]

Rao, J. N. K., and Wu, C. (2009), “Empirical Likelihood Methods,” in *Handbook of Statistics*, Vol. 29B, *Sample Surveys: Inference and Analysis*, eds. D. Pfeffermann and C.R. Rao, Amsterdam: Elsevier, pp. 189–207. [1502]

Sampford, M. R. (1967), “On Sampling Without Replacement With Unequal Probabilities of Selection,” *Biometrika*, 54, 499–513. [1499]

Skinner, C. J. (1991), “On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys,” *Journal of the American Statistical Association*, 86, 779–784. [1494]

Skinner, C. J., and Rao, J. N. K. (1996), “Estimation in Dual Frame Surveys With Complex Designs,” *Journal of the American Statistical Association*, 91, 349–356. [1494-1496,1499]

Wu, C. (2004), “Some Algorithmic Aspects of the Empirical Likelihood Method in Survey Sampling,” *Statistica Sinica*, 14, 1057–1067. [1499]

——— (2005), “Algorithms and R Codes for the Pseudo Empirical Likelihood Methods in Survey Sampling,” *Survey Methodology*, 31, 239–243. [1499]

Wu, C., and Rao, J. N. K. (2006), “Pseudo-Empirical Likelihood Ratio Confidence Intervals for Complex Surveys,” *The Canadian Journal of Statistics*, 34, 359–375. [1494,1495,1498,1502]

——— (2010), “Bootstrap Procedures for the Pseudo Empirical Likelihood Method in Sample Surveys,” *Statistics and Probability Letters*, 80, 1472–1478. [1498]