Calculating Estimates for Multiple-frame Surveys with More than Two Frames

Amber Tomas and Qiannan Yin

April 13, 2016

1 Introduction

There are now a number of different estimators for calculating estimates from multiple-frame surveys. However, most available documentation and code only applies to the case of two frames. In this document we describe how to compute estimates for the more general $Q$-frame case, and provide pseudo-code. This pseudo-code forms the basis for an R package currently in development by the authors.

2 Notation and Assumptions

Our notation largely follows that used in Lohr and Rao [2006]. Suppose there are $Q$ frames, denoted $A_1, \ldots, A_Q$. Let $F$ denote the index set of frames, so $F = 1, 2, \ldots, Q$. Assume that the union of the frames covers the population of interest. The population domains are defined by the subsets of $F$. For $k \subseteq F$, the domain defined by $D_k$ is

$$D_k = \left( \bigcap_{q \in D_k} A_q \right) \cap \left( \bigcap_{q \notin D_k} A_q^c \right),$$

where $c$ denotes complement. We use $d$ to denote the number of non-empty domains, and $q \in D_k$ to denote the set of frames that cover domain $D_k$.

Let $N^{(q)}$ be the number of population units on frame $A_q$, and let $N_k$ be the number of population units in domain $D_k$. Then $\sum_{k=1}^d N_k = N$, the population size. We use $S_q$ to denote the set of units sampled from frame $A_q$, $Y^{(q)}$ to denote the population total of units on frame $q$, and $Y_k$ to denote the population total of domain $D_k$.

The algorithms presented in this paper assume that only a small amount of information is available. In particular, we assume:

- the domain membership is known for all sampled units;
- the sampling weights from each frame are available;
- the value of the response variable is known for all sampled units from each frame.

We use $w^{(q)}_i$ to denote the weight of the $i$th unit on frame $A_q$, and $y^{(q)}_i$ to denote the value of the response variable for this unit. Often there may be additional information available, such as replicate weights and auxiliary variables. However, the methods presented in this paper can be used without this additional information. This is useful for practitioners who want to calculate estimates from publicly available data which often comes only with the values of the response variables and their sampling weights.

Unless otherwise stated, we assume that the Horvitz-Thompson estimator is used to estimate population totals. Hence for $k = 1, \ldots, d$, the estimate of $Y_k$ using the sample from frame $A_q$ is

$$\hat{Y}_k^{(q)} = \sum_{i \in S_q} w^{(q)}_i \delta_i(k) y_i$$
where

\[ \delta_i(k) = \begin{cases} 
  1 & \text{if } i \in D_k \\
  0 & \text{otherwise} 
\end{cases} \]

Similarly, the estimate of the population size in domain \( k \), \( N_k \), based on the sample from frame \( q \) is

\[ \hat{N}_k^{(q)} = \sum_{i \in S_q} w_i^{(q)} \delta_i(k). \]

The remaining sections give pseudo-code that can be used to calculate the following estimators:

- the Hartley estimator
- the Fuller-Burmeister estimator
- the Pseudo Maximum Likelihood estimator
- the Pseudo Empirical Likelihood estimator (not incorporating auxiliary information)
- the Unique Linkage estimator
- the Multiplicity estimator
- the Single-frame estimator.

We use these estimators as they have been most referenced in the literature or used in practice. Our pseudo-code assumes that the necessary variances can be calculated, and the final section discusses how these variances can be estimated if no additional design information is available.

## 3 Hartley Estimator

The optimal linear estimator proposed by H19 [1962] and generalized to Q-frames by Lohr and Rao [2006] seeks to find the linear combination of separate frame estimates in each domain which will minimize the variance of the total population estimator. It requires calculating the covariance matrix of the domain estimates from each frame, and we assume that the practitioner can do this using the best variance approximation methods available to them (for example, using replicate weights). Methods for calculating the estimated covariance matrix if no information is available other than the \( w_i^{(q)} \) and \( y_i \) are given in Section 10.

Below is pseudo-code that can be used to calculate the Hartley estimator.

1. Calculate the \( d \)-vectors

\[ \hat{Y}_q = (\hat{Y}_1^{(q)}, \ldots, \hat{Y}_d^{(q)})^T, \]

and

\[ \hat{N}_q = (\hat{N}_1^{(q)}, \ldots, \hat{N}_d^{(q)})^T, \quad q = 1, \ldots, Q. \]

Note that for \( q \not\in D_k \), i.e. if frame \( q \) does not cover domain \( D_k \), then

\[ \hat{Y}_k^{(q)} = \hat{N}_k^{(q)} = 0. \]

2. Let \( C_q \) denote the estimated covariance matrix of \( \hat{Y}_q \).

   (a) Construct \( C_q \), \( q = 1, \ldots, Q \) using one of the methods in Section 10 or an alternative.
   (b) Define the \( d \times d \) diagonal matrix \( \Gamma_q \) to have \( k \)th diagonal entry 1 if \( q \in k \) and 0 otherwise, and let \( 1_d \) be a \( d \)-vector of 1s.

   Calculate

\[ E_q = \Gamma_q C_q \Gamma_q, \quad q = 1, \ldots, Q. \]
(c) Calculate the \((dQ + 1) \times dQ\) block matrix
\[
A_{i,j} = \begin{cases} 
\Gamma_i \left( \sum_{q=1}^{Q} \Gamma_q \right)^{-1} \left( I - \sum_{q=1}^{Q} \Gamma_q \right) E_i & \text{if } i = j \\
\Gamma_i \left( \sum_{q=1}^{Q} \Gamma_q \right)^{-1} E_j & \text{if } i \neq j, i \leq Q \\
\Gamma_j & \text{if } i = Q + 1
\end{cases}
\]
for \(i = 1, \ldots, Q + 1, \ j = 1, \ldots, Q\).

(d) Calculate the \(dQ \times 1\) vector
\[
\theta = (\theta_1^T, \ldots, \theta_{Q}^T)^T
\]
which is the solution to
\[
A \theta = [0_{dQ}^T \ 1_d^T]^T
\]
The set of linear equations (1) has \(d(Q + 1)\) equations in \(dQ\) unknowns, but is of rank \(dQ\). We solve this system of linear equations using the generalized inverse of \(A\).

(e) The final estimate is then given by
\[
\hat{Y}_H(\theta) = \sum_{q=1}^{Q} \theta_q \Gamma_q \hat{Y}_q^T.
\]
Equivalently, so-called “modified” weights can be calculated. These are the weights that can be used for calculating estimates if the \(q\) samples are concatenated together and treated as a single sample. Then the modified weights can be calculated as follows:
\[
\tilde{w}_i^{(q)} = w_i^{(q)} \theta_q[k], \ q = 1, \ldots, Q, \ i \in S_q,
\]
where \(\theta_q[k]\) is the \(k\)-th element of \(\theta_q\). Then the final estimate can be calculated as
\[
\hat{Y}_H(\theta) = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i^{(q)}.
\]

4 Fuller-Burmeister estimator

The Fuller-Burmeister estimator [FB1, 1972] is similar to the Hartley estimator but incorporates information about the estimates \(\hat{N}(q)\) as well as \(\hat{Y}(q)\) to find an optimal linear combination of domain estimates. It was also generalized to Q-frames by Lohr and Rao [2006], and here we present pseudo-code that can be used to calculate it.

1. Estimate the \(2d \times 2d\) covariance matrix
\[
D_q = \text{cov} \left[ \begin{bmatrix} \hat{Y}_q \\ \hat{N}_q \end{bmatrix} \right] = \begin{bmatrix} C_q & D_{q12} \\ D_{q12}^T & D_{q22} \end{bmatrix}
\]

2. Let \(G_q\) be the \(2d \times 2d\) matrix \(\text{diag}(\Gamma_q, \Gamma_q)\), where \(\Gamma_q\) is as defined in Section 3. Calculate
\[
E_q = G_q D_q G_q, \ q = 1, \ldots, Q
\]

3. Construct the \(2d(Q + 1) \times 2dQ\) block diagonal matrix
\[
B_{i,j} = \begin{cases} 
G_i \left( \sum_{q=1}^{Q} G_q \right)^{-1} \left( I - \sum_{q=1}^{Q} G_q \right) E_i & \text{if } i = j \\
G_i \left( \sum_{q=1}^{Q} G_q \right)^{-1} E_j & \text{if } i \neq j, i \leq Q \\
G_j & \text{if } i = Q + 1
\end{cases}
\]
for \(i = 1, \ldots, Q + 1, \ j = 1, \ldots, Q\).
4. Compute the $2dQ$-vector $\beta = (\beta_1^T, \ldots, \beta_Q^T)^T$, as the solution to

$$B\beta = [0_{2dQ}^T 1_d^T 0_d^T]^T$$

(2)

The set of linear equations (2) has $2d(Q+1)$ equations in $2dQ$ unknowns, but is of rank $2dQ$. We solve this system of linear equations using the generalized inverse of $B$.

5. Then

$$\hat{Y}_{FB}(\beta) = \sum_{q=1}^Q \beta_q^T G_q [\hat{Y}^T_q \hat{N}_q^T]^T.$$  

5 Pseudo Maximum Likelihood Estimator

Pseudo maximum likelihood estimation for dual frame surveys using simple random sampling was first proposed by Skinner [1991]. This was extended to $Q$-frame complex surveys by Lohr and Rao [2006].

1. Let

$$f^{(q)} = \frac{n^{(q)}}{N(q)d^{(q)}}, \quad q = 1, \ldots, Q,$$

where $d^{(q)}$ is the design effect for estimating $Y$ from frame $q$. For simple random sampling the design effect is equal to 1, whereas for complex samples typically $d^{(q)}$ is greater than 1. Lohr and Rao [2006] discuss methods for how to choose a design effect if it is not approximately the same for all variables.

If the frame sizes are not known, we use

$$f^{(q)} = \frac{n^{(q)}}{N(q)d^{(q)}}, \quad q = 1, \ldots, Q.$$  

2. Calculate

$$\hat{Y}_k = \frac{\sum_{q \in D_k} \hat{f}^{(q)} \hat{Y}_k^{(q)}}{\sum_{q \in D_k} \hat{f}^{(q)} \hat{N}_k^{(q)}}, \quad k = 1, \ldots, d.$$  

3. Define $M$ to be the $d \times Q$ matrix with $k,q$th entry

$$M_{k,q} = \begin{cases} 1 & \text{if } q \in D_k, \\ 0 & \text{otherwise}. \end{cases}$$

, for $q = 1, \ldots, Q, \quad k = 1, \ldots, d$.

4. Let $\hat{H}$ be the $d \times Q$ matrix with $k,q$th entry

$$\hat{H}_{k,q} = \begin{cases} \hat{N}_k^{(q)} & \text{if } q \in D_k, \\ 0 & \text{otherwise}. \end{cases}$$

, for $q = 1, \ldots, Q, \quad k = 1, \ldots, d$.

5. The aim of this step is to find the $d$-vector of domain population sizes $\hat{N}$, which is the solution to

$$\begin{bmatrix} (I - MM^+)(\text{diag}\hat{N})^{-1}\hat{H}f \\ M^T\hat{N} - h \end{bmatrix} = 0_{Q+d},$$

(3)

where $M^+$ is the Moore-Penrose generalized inverse of $M$, and $h = (N^{(1)}, N^{(2)}, \ldots, N^{(Q)})^T$. $M$ is nonsingular only if we know the true values of $N$, in which case this step should be skipped and the true values used in place of the estimated frame sizes.

If $N$ is not known, the following method can be used to find an approximate solution to (3) [Lohr and Rao, 2006]: This relationship can be used to iteratively calculate $\hat{N}$ using the following method:

i) Let $\hat{N}_0$ be a vector of initial estimates for $N$. Two reasonable options are

(a) If frame $q$ has the largest sample size in domain $D_k$, then let $\hat{N}_{k0} = \hat{N}_k^{(q)}$;
Let $\tilde{N}_{k0}$ equal the average of $\hat{N}_k^{(q)}$, $q = 1, \ldots, Q$.

ii) For $l = 1, \ldots$ calculate $\tilde{N}_{l+1}$, which is the solution to the linear set of equations

$$\begin{bmatrix}
(I - MM^+)(\text{diag} \tilde{N}_l)^{-1}(\text{diag} Mf) \\
M^T
\end{bmatrix}
\tilde{N}_{l+1}
= \begin{bmatrix}
(I - MM^+)(\text{diag} \tilde{N}_l)^{-1} \hat{H}_f \\
h
\end{bmatrix}.
$$

iii) If the difference between $N_{l+1}$ and $N_l$ is small enough, let $\hat{N}_l = N_{l+1}$ and stop iterating.

In our studies we used option (b) for $\tilde{N}_{k0}$, and a stopping distance of 0.001.

6. For unit $i$ on frame $A_q$ in domain $D_k$, calculate the adjusted weight

$$\tilde{w}_{iq}^{(q)} = \frac{w_i^{(q)} N_k f^{(q)}}{\sum_{p \in D_k} f^{(p)} \tilde{N}_k^{(p)}}, \quad q = 1, \ldots, Q.$$

7. Calculate the final estimate

$$\hat{Y}_{PML} = \sum_{q=1}^Q \sum_{i \in S_q} \tilde{w}_{iq}^{(q)} y_i^{(q)}.$$

### 6 Pseudo-Empirical Likelihood Estimator

Pseudo-empirical likelihood (PEL) estimation for multiple frame surveys was introduced by Rao and Wu [2010]. They suggested two possible approaches:

1. PEL methods for dual-frame surveys based on post-stratified samples;
2. A multiplicity-based PEL approach which requires knowledge only of the multiplicity of each sampled observation and not the domain membership information, and which is easily extended to more than two frames.

In the absence of auxiliary variables, the multiplicity approach simplifies to that of Mecatti [2007], which we describe in Section 8. Since we do not assume the presence of any auxiliary variables for this paper, we only present the PEL post-stratified estimator. Rao and Wu [2010] presented the post-stratified estimator only for the dual frame case, and noted that it could be extended to the more general $Q$-frame case but that the general case was not presented due to notational difficulties. We agree that the notational complexities are real, and for the same reason here we present pseudo-code that can be used to calculate the PEL estimator for both the 2- and 3-frame case, and indicate how it can be extended further if required.

1. Construct the vectors of normalized weights\(^1\) $\tilde{d}_k^{(q)}$: a vector of length $n_k^{(q)}$ that has $i$th element

$$\tilde{d}_{ki}^{(q)} = \frac{w_{i}^{(q)}}{\sum_{i \in S_q \cap D_k} w_{i}^{(q)}}, \quad \forall \ i \in S_q \cap D_k.$$

There should be $d \times Q$ such vectors: one for every frame-domain combination.

2. The next step is to determine values for the constants $\eta_{kq}$, $k = 1, \ldots, d$, $q = 1, \ldots, Q$. The constant $\eta_{kq}$ can be thought of as the “weight” given to the estimate from frame $q$ in domain $k$. Rao and Wu [2010] propose two possibilities:

\(^1\)Rao and Wu [2010] suggest using the inverse selection probabilities in place of weights when available.
A. Calculate 
\[ \hat{\eta}_{k,q} = \frac{\hat{V}(\hat{Y}_k^{(q)})}{\sum_{l \in D_k} \hat{V}(\hat{Y}_l^{(l)})}, \quad k = 1, \ldots, d, \ q = 1, \ldots, Q, \]
where \( \hat{Y}_k^{(q)} \) is the ratio estimator
\[ \hat{Y}_k^{(q)} = \frac{\sum_{i \in S_q \cap D_k} w_i^{(q)} y_i}{\sum_{i \in S_q \cap D_k} w_i^{(q)}}, \]

B. Calculate \( \hat{\eta}_{k,q}^{(q)} = \frac{\hat{V}(\hat{N}_k^{(q)})}{\sum_{l \in D_k} \hat{V}(\hat{N}_l^{(l)})}, \quad k = 1, \ldots, d, \ q = 1, \ldots, Q. \)

3. Calculate the “weights”
\[ W_k^{(q)} = \frac{\hat{N}_k^{(q)}}{N^{(q)}} \times \hat{\eta}_{k,q}^{(q)}, \quad k = 1, \ldots, d, \ q = 1, \ldots, Q. \]

4. We break this step up and present the pseudo-code separately for \( Q = 2 \) and then \( Q = 3. \) At the end of this section an explanation is given for how to extend the method to more than three frames.

A. Suppose there are only two frames, A and B. Let domain \( D_a \) include units only on frame A, domain \( D_{ab} \) include units on both frame A and frame B, and domain \( D_b \) include units only on frame B. Construct the following matrices and vectors:

\[ n = (n_a^{(A)}, n_{ab}^{(A)}, n_{ab}^{(B)}, n_b^{(B)})^T, \quad (4 \times 1 \text{ vector}) \]
\[ x_a^{(A)} = 0_{n_a^{(A)} \times 1} \]
\[ x_{ab}^{(A)} = \frac{y_i}{W_{ab}^{(A)}}, \quad i \in S_A \cap D_{ab}, \quad (n_{ab}^{(A)} \times 1 \text{ vector}) \]
\[ x_{ab}^{(B)} = \frac{y_i}{W_{ab}^{(B)}}, \quad i \in S_B \cap D_{ab}, \quad (n_{ab}^{(B)} \times 1 \text{ vector}) \]
\[ x_b^{(B)} = 0_{n_b^{(B)} \times 1} \]
\[ x = (x_a^{(A)}, x_{ab}^{(A)}, x_{ab}^{(B)}, x_b^{(B)}), \quad (n \times 1 \text{ vector}) \]
\[ d = (\tilde{d}_a^{(A)}, \tilde{d}_{ab}^{(A)}, \tilde{d}_{ab}^{(B)}, \tilde{d}_b^{(B)}), \quad (n \times 1 \text{ vector}) \]
\[ X = 0, \quad \text{scalar} \]
\[ W = (W_a^{(A)}, W_{ab}^{(A)}, W_{ab}^{(B)}, W_b^{(B)}), \quad (4 \times 1 \text{ vector}) \]

B. Suppose there are only three frames, A, B, and C. Let \( D_{ab} \) denote the domain where only frames A and B overlap, and \( D_{abc} \) denote the domain where all three frames overlap. Other domains are denoted similarly. Construct the following matrices and vectors:

\[ \text{See equation (2.8) in Rao and Wu [2010] for an adjustment they suggest in the 2-frame case.} \]
and this paper the problem is formulated as follows: the calculation of the PEL estimate for the case of stratified sampling is given in Wu [2005]. In Rao and Wu [2010] mention that the multiple-frame PEL estimator can be considered as a case of the PEL frames Explanation of Calculations for PEL Estimator in the case of more than two frames

5. Use code in Appendix A3. of Wu [2005] to solve for $\hat{\theta}$, a vector of length $n$.

6. Calculate the estimate

$$\hat{Y}_{PEL} = \sum_k \sum_{q \in D_k} W_k^{(q)} \sum_{i \in S_i \cap D_k} \hat{p}_k^{(q)} y_i.$$  

**Explanation of Calculations for PEL Estimator in the case of more than two frames**

Rao and Wu [2010] mention that the multiple-frame PEL estimator can be considered as a case of the PEL estimator under stratified sampling, where each domain/frame combination can be considered as a separate stratum. The calculation of the PEL estimate for the case of stratified sampling is given in Wu [2005]. In this paper the problem is formulated as follows:

Maximize

$$l(p_1, \ldots, p_H) = n^* \sum_{h=1}^H W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}),$$

subject to the constraints

$$p_{hi} > 0, \quad \sum_{i \in s_h} p_{hi} = 1, \quad h = 1, \ldots, H$$

and

$$\sum_h W_h \sum_{i \in s_h} p_{hi} x_{hi} = \mathbf{x},$$  \hfill (4)
where here $x_{hi}$ is a vector of auxiliary variables, $X$ the vector of known population totals for the auxiliary variables, and $s_h$ is the set of sampled units from stratum $h$.

In Rao and Wu (2010) the constrained maximization problem is presented as:

Maximize

$$l(p_a, p_{ab}, p_{ba}, p_b) = n^* \sum_h W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}), \quad h \in \{a, ab, ba, b\},$$

subject to the constraints

$$\sum_{i \in s_h} p_{hi} = 1, \quad h = a, ab, ba, b,$$

and

$$\sum_{i \in s_{ab}} p_{abi} y_i = \sum_{j \in s_{ba}} p_{baj} y_j.$$  \hspace{1cm} (5)

Note that they do not specify that the $p_{hi}$ must be greater than zero, although we assume that this constraint is intended. The constraint (5) can be reexpressed as

$$W_{ab} \sum_{i \in s_{ab}} p_{abi} y_i - \sum_{j \in s_{ba}} p_{baj} y_j = 0, \text{ or equivalently}$$

$$W_{ab} \sum_{i \in s_{ab}} p_{abi} y_i = \sum_{j \in s_{ba}} p_{baj} y_j.$$  \hspace{1cm} (6)

This can be formulated similarly to (6) but using one auxiliary variable for each equality in (7).

The Unique Linkage (UL) Estimator has been used in multiple-frame surveys such as SESTAT [NCSES, b,a]. The idea is that each sampled unit is linked to just one frame, and information on that unit from the other frames is ignored for estimation purposes.

1. Determine an order for the $Q$ frames. Denote the rank of frame $q$ in this order by $R(q)$.

One method for determining the order is as follows:

(a) Calculate $\hat{V}(\hat{Y}^{(q)})$, for $q = 1, \ldots, Q$.

(b) Order the surveys by increasing value of $\hat{V}(\hat{Y}^{(q)})$. 

7 Unique Linkage Estimator

The Unique Linkage (UL) Estimator has been used in multiple-frame surveys such as SESTAT [NCSES, b,a]. The idea is that each sampled unit is linked to just one frame, and information on that unit from the other frames is ignored for estimation purposes.

1. Determine an order for the $Q$ frames. Denote the rank of frame $q$ in this order by $R(q)$.

One method for determining the order is as follows:

(a) Calculate $\hat{V}(\hat{Y}^{(q)})$, for $q = 1, \ldots, Q$.

(b) Order the surveys by increasing value of $\hat{V}(\hat{Y}^{(q)})$. 

8
2. Let $M[i]$ denote the set of frames which contain unit $i$. Calculate weight modification factors

$$m_i^{(q)} = \begin{cases} 1 & \text{if } R(q) = \min_{k \in M[i]} R(k), \quad q = 1, \ldots, Q, \; i \in S_q \\ 0 & \text{otherwise} \end{cases}$$

3. Calculate the modified weights

$$\tilde{w}_i^{(q)} = m_i^{(q)} w_i^{(q)}, \quad q = 1, \ldots, Q, \; i \in S_q.$$  

4. The Unique Linkage estimator is given by

$$\hat{Y}_{UL} = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i.$$  

8 The Multiplicity Estimator

Compared to previous estimators we’ve considered, the multiplicity estimator requires one additional assumption: that the number of frames to which every sampled unit belongs is known. The number of frames on which a unit is listed is referred to as the multiplicity of that unit. The multiplicity estimator for multiple-frame surveys was first proposed by Mecatti [2007].

1. Let $m_i$ be the multiplicity of unit $i$, i.e. the number of frames that include unit $i$.

2. Calculate the modified weights of the sampled units

$$\tilde{w}_i^{(q)} = w_i^{(q)} m_i^{-1}, \quad q = 1, \ldots, Q; \; i \in S_q.$$  

3. Then the Mecatti estimator is given by

$$\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i^{(q)}.$$  

9 The Single-frame Estimator

This estimation method requires the following additional assumption: that the records of every sampled unit can be identified on all the frames that include it. This is stronger than the assumption required for the Mecatti estimator. The single-frame estimator was first proposed for dual-frame surveys by Kalton and Anderson [1986] and then extended to the Q-frame case by Lohr and Rao [2006].

1. Let $w_i^{(q)}$ be the weight on frame $A_q$ of population unit $i$. Calculate the modified weights of the sampled units

$$\tilde{w}_i^{(q)} = \left( \sum_{j=1}^{Q} \frac{1}{w_i^{(j)}} \right)^{-1}, \; q = 1, \ldots, Q; \; i \in S_q.$$  

2. Then the single-frame estimator is

$$\hat{Y}_{SF} = \sum_{q=1}^{Q} \sum_{i \in S_q} \tilde{w}_i^{(q)} y_i^{(q)}.$$  

9
10 Variance Estimation

As stated in the introduction, we assume only that the frame and domain membership is known, along with the frame sampling weights and value of the response variable. Most of the estimators presented in previous sections require estimates of the variance of estimates of the population mean or total. Typically this would be done using replicate weights or Jacknife methods, both of which assume knowledge of the sampling design and may require knowledge of design variables not on the available data set. Therefore, in this section we present several options for how the variances can be estimated in the absence of additional information.

10.1 Variance Estimation for the Pseudo-Empirical Likelihood Estimator

Step 2. of the Pseudo-Empirical Likelihood Estimator presented in section 6 uses variance estimates of the estimated domain means from each frame. The variance of the Hajek estimators of the domain mean can be written as follows:

$$V(\hat{\bar{Y}}_{kH}^{(q)}) = V\left[\frac{\hat{Y}_{k}^{(q)}}{N_{k}^{(q)}}\right]$$

$$= V\left[\frac{\sum_{i\in S_{k}^{(q)}} w_{i}^{(q)} y_{i}}{\sum_{i\in S_{k}^{(q)}} w_{i}^{(q)}}\right]$$

$$= V\left[\frac{\sum_{i\in S_{k}^{(q)}} I_{ki} w_{i}^{(q)} y_{i}}{\sum_{i\in S_{k}^{(q)}} I_{ki} w_{i}^{(q)}}\right],$$

where

$$I_{ki} = \begin{cases} 1 & \text{if unit } i \text{ is in domain } D_{k} \\ 0 & \text{otherwise} \end{cases}.$$  \hspace{1cm} (8)

The required variances are therefore all variances of some ratio estimator, for some frame \( q = 1, \ldots, Q \) and domain \( D_{k}, k = 1, \ldots, d \).

If SRS is used to select the sample from frame \( q \), then the weights cancel out and the ratio estimator from (8) can be written as

$$\tilde{B}_{k}^{(q)} = \frac{\sum_{i\in S_{q}^{(q)}} I_{ki} y_{i}}{\sum_{i\in S_{q}} I_{ki}}.$$  \hspace{1cm} (9)

The variance of \( \tilde{B}_{k}^{(q)} \) can be estimated by (see for example Lohr [2010, p. 126])

$$V(\tilde{B}_{k}^{(q)}) = \left(1 - \frac{n^{(q)}}{N^{(q)}}\right) \frac{1}{n^{(q)} - 1} \sum_{i\in S_{q}^{(q)}} (y_{i} I_{ki} - \tilde{B}_{k}^{(q)} I_{ki})^{2} \frac{1}{n^{(q)} I_{ki}^{2}}$$

$$= \left(1 - \frac{n^{(q)}}{N^{(q)}}\right) \frac{1}{n^{(q)} - 1} \sum_{i\in S_{q}^{(q)} \cap D_{k}} (y_{i} - \tilde{B}_{k}^{(q)})^{2} / n^{(q)}.$$  \hspace{1cm} (9)

If SRS is not used, we use the following weighted variance formula:

$$V(\tilde{B}_{k}^{(q)}) = \left(1 - \frac{n^{(q)}}{N^{(q)}}\right) \frac{1}{n^{(q)} - 1} \sum_{i\in S_{q}^{(q)} \cap D_{k}} w_{i}^{(q)} (y_{i} - \tilde{B}_{k}^{(q)})^{2} \frac{1}{\sum_{i\in S_{q}^{(q)} \cap D_{k}} w_{i}^{(q)}},$$

which simplifies to (9) when the weights are equal for all units sampled from frame \( q \) in domain \( k \). If the effective sample size for frame \( q \) is known, this should be used in place of \( n^{(q)} \).

10.2 Variance and Covariance Estimation for the Hartley and Fuller-Burmeister Estimators

The generalized Hartley estimator and Fuller-Burmeister estimator are functions of \( C_{q} \), the covariance matrix of \( \hat{Y}_{d}^{(q)} \), for all domains \( d \) covered by frame \( q \), for \( q = 1, \ldots, Q \). Firstly we derive formulas for the variances in this matrix, and then the covariances.
10.2.1 Variances

Let

\[ u_i = \begin{cases} y_i & i \in U_q \cap D_{k1} \\ 0 & \text{otherwise,} \end{cases} \]

where \( U_q \cap D_{k1} \) is the set of all population units from frame \( q \) in domain \( k1 \). Then

\[ Y_{k1}^{(q)} = \sum_{i \in U_q} u_i. \]

Assuming that \( Y_{k1}^{(q)} \) is estimated by the Horvitz-Thompson estimator

\[ \hat{Y}_{k1}^{(q)} = \sum_{i \in S(q)} \pi_i^{-1(q)} u_i, \]

the variance of this estimator is given by

\[ V(\hat{Y}_{k1}^{(q)}) = \sum_{i \in U_q} \left( \frac{1 - \pi_i^{(q)}}{\pi_i^{(q)}} \right) u_i^2 + \sum_{i \in U_q} \sum_{j \neq i} \left( \frac{\pi_i^{(q)} - \pi_i^{(q)} \pi_j^{(q)}}{\pi_i^{(q)} \pi_j^{(q)}} \right) u_i u_j. \]

We assume that only the sampling weights are known, and that \( w_i^{(q)} \approx \pi_i^{-1(q)} \) for all units \( i \) on frame \( q \), \( q = 1, \ldots, Q \). We consider the following three approximations to the variance (11):

1. \[ \hat{V}(\hat{Y}_{k1}^{(q)}) = (N^{(q)})^2 \left( 1 - \frac{n^{(q)}}{N^{(q)}} \right) \frac{s_u^2}{n^{(q)}}, \]

where

\[ s_u^2 = \frac{1}{n^{(q)} - 1} \sum_{i \in S_q} (u_i - \bar{u})^2. \]

This is the unbiased estimator of (10) in the case that SRSWOR was used to select the sample from frame \( q \).

2. \[ \hat{V}(\hat{Y}_{k1}^{(q)}) = \sum_{i \in S_q} w_i^{(q)} (w_i^{(q)} - 1) y_i^2. \]

This is obtained by assuming that \( \pi_{ij} \approx \pi_i \pi_j \forall i, j \), i.e. that units are sampled independently, in which case we can write (11) as

\[ \hat{V}(\hat{Y}_{k1}^{(q)}) \approx \sum_{i \in U_q} \left( \frac{1 - \pi_i^{(q)}}{\pi_i^{(q)}} \right) y_i^2. \]

An unbiased estimate of this is given by

\[ \hat{V}(\hat{Y}_{k1}^{(q)}) = \sum_{i \in S_q} \pi_i^{-1(q)} \left( \frac{1 - \pi_i^{(q)}}{\pi_i^{(q)}} \right) y_i^2 = \sum_{i \in S_q} w_i^{(q)} (w_i^{(q)} - 1) y_i^2. \]

We would expect this estimator to do less well when the sampling fraction is large, in which case the assumption that the second term in the variance is close to zero is unlikely to hold.
We again consider three possibilities, based on different assumptions:

3. 
\[
\hat{V}(\hat{Y}_{k1}^{(q)}) = \frac{n^{(q)}}{n^{(q)}-1} \sum_{i \in S_q} (1 - \pi_i^{(q)}) (u_i \pi_i^{-1} - A_s^{(q)})^2,
\]

where
\[
A_s^{(q)} = \sum_{i \in S_q} \frac{1 - \pi_i^{(q)}}{\sum_{j \in S_q} (1 - \pi_j^{(q)})} u_i \pi_i^{-1}.
\]

Henderson [2006] compared nine estimators of the variance of the Horvitz-Thompson estimator (11). The estimator (12), first derived by Hajek (1964), seemed to perform well in many cases. However, it does take significantly longer to compute than estimates 1. and 2.

10.2.2 Covariances

We assume that samples from different frames are independent, in which case Cov(\(\hat{Y}_{k1}^{(q)}, \hat{Y}_{k2}^{(q)}\)) = 0 for all \(q1 \neq q2\). Therefore, we only need to consider the covariance between domain estimates for estimates calculated from the same frame.

Let
\[
v_i = \begin{cases} y_i & i \in U_q \cap D_{k2} \\ 0 & \text{otherwise}, \end{cases}
\]

An unbiased estimator of the covariance between two Horvitz-Thompson estimators
\[
\hat{Y}_{k1}^{(q)} = \sum_{i \in S_q} w_i u_i, \quad \text{and} \quad \hat{Y}_{k2}^{(q)} = \sum_{i \in S_q} w_i v_i,
\]
is [Särndal et al., 1992]

\[
\widehat{\text{Cov}}(\hat{Y}_{k1}^{(q)}, \hat{Y}_{k2}^{(q)}) = \sum_{i \in S_q} \sum_{j \in S_q} \pi_{ij}^{(q)} \frac{\pi_i^{(q)} - \pi_j^{(q)}}{\pi_{ij}^{(q)}} \frac{u_i}{\pi_i^{(q)}} \frac{v_j}{\pi_j^{(q)}}
\]
\[
= \sum_{i \in S_q} (1 - \pi_i^{(q)}) \frac{u_i}{\pi_i^{(q)}} \frac{v_i}{\pi_i^{(q)}} + \sum_{i \in S_q} \sum_{j \neq i} \pi_{ij}^{(q)} \frac{u_i}{\pi_i^{(q)}} \frac{v_j}{\pi_j^{(q)}}.
\]

Since it is assumed that the sampling weights of only the sampled units are known, we need to approximate (13). We again consider three possibilities, based on different assumptions:

1. 
\[
\widehat{\text{Cov}}(\hat{Y}_{k1}^{(q)}, \hat{Y}_{k2}^{(q)}) = \frac{N^{(q)}(N^{(q)} - n^{(q)})}{(n^{(q)})^2} \sum_{i \in S_q} u_i v_i + \sum_{i \in S_q} \sum_{j \neq i} \left( \frac{(N^{(q)})^2}{(n^{(q)})^2} - \frac{N^{(q)} N^{(q)} - 1}{n^{(q)} n^{(q)} - 1} \right) u_i v_j.
\]

This is obtained from (13) under the assumption that SRSWOR was used, in which case we know \(\pi_i^{(q)} = n^{(q)}/N^{(q)}\) and \(\pi_{ij}^{(q)} = n^{(q)}/N^{(q)}(n^{(q)} - 1)/(N^{(q)} - 1)\), \(i \neq j\).

2. 
\[
\widehat{\text{Cov}}(\hat{Y}_{k1}^{(q)}, \hat{Y}_{k2}^{(q)}) = \sum_{i \in S_q} w_i^{(q)} (w_i^{(q)} - 1) u_i v_i.
\]

This is obtained from (13) under the assumption that \(\pi_{ij}^{(q)} \approx \pi_i^{(q)} \pi_j^{(q)} \forall i, j\). In this case the second term is approximately equal to zero.

3. 
\[
\widehat{\text{Cov}}(\hat{Y}_{k1}^{(q)}, \hat{Y}_{k2}^{(q)}) = \sum_{i \in S_q} (1 - \pi_i^{(q)}) \frac{u_i}{\pi_i^{(q)}} \frac{v_i}{\pi_i^{(q)}} + \sum_{i \in S_q} \sum_{j \neq i} \left( 1 - \frac{1 - \pi_i^{(q)}(1 - \pi_j^{(q)})}{\sum_{l \in S_q} (1 - \pi_l^{(q)})} \right) \frac{u_i}{\pi_i^{(q)}} \frac{v_j}{\pi_j^{(q)}}.
\]
This is obtained based on substituting the following approximation [Hájek, 1964] for the joint selection probabilities:

\[ \pi^{(q)}_{ij} = \pi^{(q)}_i \pi^{(q)}_j \left[ \frac{1 - (1 - \pi^{(q)}_i)(1 - \pi^{(q)}_j)}{\sum_{l \in U_q} \pi^{(q)}_l (1 - \pi^{(q)}_k)} \right] \]

and approximating the denominator of this term by

\[ \sum_{l \in S_q} (1 - \pi^{(q)}_l). \]

Similar to the methods for estimating variances, method 2. is the most computationally efficient closely followed by method 1. The third method requires significantly more computation time. We’d expect method 1 to be most accurate if there is small weight variation between units on the same frame. If this is not the case then method 2. may give more accurate results, particularly if the sampling fraction is small, in which case the independence assumptions are more reasonable.

References


