VARIANCE ESTIMATION FOR 2000 CENSUS COVERAGE ESTIMATES

Jae Kwang Kim, Westat; Alfredo Navarro, Bureau of Census; Wayne Fuller, Iowa State University
Jae Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Variance estimation, double sampling, double expansion estimator

1. Introduction

Two-phase sampling, also known as double sampling, can be a cost-effective technique in large-scale surveys. By first selecting a large sample, observing cheap auxiliary variables and then by properly incorporating the auxiliary variables into the second phase sampling design, we can produce estimators with smaller variances than those based on a single phase sampling design for the same cost. In one of the common procedures of two-phase sampling, the second phase sample is selected using stratified sampling where the strata are created after observing the first phase.

Rao (1973) and Cochran (1977) give formulas for variance estimation when the first phase is a simple random sample and the second phase is stratified random sampling. Kott (1990) derived a formula for variance estimation when the first phase is a stratified random sample and the second phase is a re-stratified random sample based on first–phase information. Rao and Shao (1992) proposed a jackknife variance estimation method in the context of hot deck imputation where the second phase strata correspond to imputation cells. Fuller (1998) proposed a replicate variance estimation method for the two-phase regression estimator.

Among the cited methods, only the Rao and Shao (1992) method and Fuller (1998) method is replication methods. One advantage of using the replication method for variance estimation is its convenience for a multi-purpose survey. That is, after we create the replication weights, we can directly apply the replication weights for any variable.

Let the parameter of interest be the population total $Y = \sum_{i=1}^{N} y_i$, where y_i is the study variables and N is assumed to be known. Suppose we have a sample with the set of indices A_1, and observe y_i on every element of the sample then

$$\hat{Y}_1 = \sum_{i\in A_1} w_i y_i , \tag{1}$$

where $w_i = [\Pr(i \in A_1)]^{-1}$, is an unbiased estimator of Y.

Now assume, instead of directly observing y_i for $i \in A_1$, we observe

$$\mathbf{x}_i = (x_{i1},\ldots,x_{iG}), \tag{2}$$

for all $i \in A_1$, where x_{ig} takes the value one if unit i belongs to the g-th group and is zero otherwise. Assume that $\sum_{g=1}^{G} x_{ig} = 1$.

Let a subsample of total size r be selected from the n first phase sample. Let A_2 be the set of indices for the second phase sample. Let

$$w_i^* = [Pr(i \in A_2 | i \in A_1)]^{-1}. \tag{3}$$

Let $n_g = \sum_{i\in A_1} x_{ig}$ be the number of first phase sample elements in group g and $r_g = \sum_{i\in A_2} x_{ig}$ be the number of second phase sample elements in group g. If the second phase sample is selected by stratified random sampling, $w_i^* = r_g^{-1} n_g$ for unit i with $x_{ig} = 1$. In this case, the groups are the second phase strata.

Given the described two phase sampling, an unbiased estimator for the total of Y is

$$\hat{Y}_d = \sum_{i\in A_2} \alpha_i y_i , \tag{4}$$

where $\alpha_i = w_i w_i^*$. Kott and Stukel (1997) called the estimator in Equation (4) the *double expansion* estimator (DEE).

Another important estimator for the total of Y is

$$\begin{aligned}\hat{Y}_r &= \sum_{g=1}^{G}\left(\sum_{i\in A_1} w_i x_{ig}\right)\frac{\sum_{i\in A_2} w_i x_{ig} y_i}{\sum_{i\in A_2} w_i x_{ig}} \\ &= \sum_{i\in A_2} \alpha_i y_i,\end{aligned} \tag{5}$$

where

$$\alpha_i' = \sum_{g=1}^{G}\left(\frac{\sum_{j\in A_1} w_j x_{jg}}{\sum_{j\in A_2} w_j x_{jg}}\right) w_i x_{ig} .$$

Kott and Stukel (1997) called the estimator in Equation (5) the *reweighted expansion* estimator (REE).

For the reweighted expansion estimator, a replication method was proposed by Rao and Shao (1992) produces consistent variance estimates. In the next subsection, we discuss the replicate method for estimating the variance of the reweighted expansion estimator. In Section 3, the replication method is extended to the double expansion estimator and multi-phase sampling. In the last section, an application used in the 2000 Census is illustrated.

2. A Replication Method for the REE

Let the replicate variance estimator for the complete sample estimator $\hat{Y}_1$ in Equation (1) be written of the form

$$\hat{V}_1 = \sum_{k=1}^{L} c_k \left(\hat{Y}_1^{(k)} - \hat{Y}_1\right)^2 , \tag{6}$$

where $\hat{Y}_1^{(k)}$ is the k-th version of $\hat{Y}_1$ based on the observations included in the k-th replicate, L is the number of replications, and c_k is a factor associated with replicate k determined by the replication method. The k-th replicate for $\hat{Y}_1$ be can be written in the form

$$\hat{Y}_1^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i , \tag{7}$$

where $w_i^{(k)}$ denotes the replicate weight for the i-th unit in the k-th replication.

Assume that the complete sample variance estimator is unbiased;

$$E\left(\hat{V}_1\right) = Var\left(\hat{Y}_1\right), \tag{8}$$

where the distribution in Equation (8) is with respect to the first phase sampling.

Rao and Shao (1992) proposed an adjusted jackknife method in the context of the hot deck imputation, where the imputation class corresponds to the second phase stratum. The proposed jackknife replicate for the reweighted expansion estimator in Equation (5) is

$$\hat{Y}_r^{(k)} = \sum_{g=1}^{G} \left(\sum_{i \in A_1} w_i^{(k)} x_{ig} \right) \frac{\sum_{i \in A_2} w_i^{(k)} x_{ig} y_i}{\sum_{i \in A_2} w_i^{(k)} x_{ig}} , \tag{9}$$

where $w_i^{(k)}$ are the full sample replicate weights of Equation (7). The replicate variance estimator can be written as

$$\hat{V}_r = \sum_{k=1}^{L} c_k \left(\hat{Y}_r^{(k)} - \hat{Y}_r\right)^2 . \tag{10}$$

Kott and Stukel (1997) suggest the use of the adjusted jackknife method defined by Equations (9) and (10) to estimate the variance of the reweighted expansion estimator.

3. Extensions to DEE and Multi-Phase Sampling

Both the REE and the DEE can be written in the form

$$\hat{Y}_2 = \hat{\mathbf{c}}_1' \hat{\mathbf{b}}_2 := \sum_{i \in A_2} \alpha_i y_i , \tag{11}$$

where $\hat{\mathbf{c}}_1$ is a vector of the estimated population characteristics estimated with the first phase sample, α_i are the coefficients that are functions of the sample but not of y, and A_2 is the set of indices for the second phase sample. The estimator $\hat{\mathbf{c}}_1$ is the realized first phase sample size in DEE and is the estimated size of the second phase strata calculated from the first phase sample for REE. In either case, we can write

$$\hat{\mathbf{c}}_1 = \sum_{i \in A} w_i q_i \mathbf{x}_i , \tag{12}$$

where $\mathbf{x}_i$ is the vector of the second phase stratum indicator functions for unit i as defined in Equation (2), $q_i = 1$ for REE and $q_i = w_i^{-1}$ for DEE. If we define $\mathbf{c}_i = q_i \mathbf{x}_i$, then $\hat{\mathbf{c}}_1$ is the estimated total of c.

The weights α_i is the α_i that minimize

$$\sum_{i \in A_2} (\alpha_i - w_i)^2 w_i^{-1} q_i , \tag{13}$$

subject to

$$\sum_{i \in A_2} \alpha_i q_i \mathbf{x}_i = \hat{\mathbf{c}}_1 , \tag{14}$$

where $\hat{\mathbf{c}}_1$ is defined in Equation (12).

Using the linearity properties of the estimator, we propose a computational procedure for variance estimation under two-phase sampling. The proposed procedure has the following features:

1. The replication weights are calculated only for the second phase sample, not for the whole first phase sample.
2. Once calculated, the replication weights can be applied to any study variables of interest.
3. The resulting replicates give the same variance estimate as the Rao-Shao method for the REE estimator.
4. The method is applicable to an extended class of regression estimators and replication methods.
5. The number of replications is the first phase sample size *n*, although methods can be developed to reduce the number of replicates.

The procedure for jackknife replication is the following:

Step 1. Delete a first phase primary sampling unit;

Step 2: Compute $\hat{\mathbf{c}}_1^{(k)}$, the first phase estimate of the total of $q_i\mathbf{x}_i$ with k-th observation deleted. That is

$$\hat{\mathbf{c}}_1^{(k)} = \sum_{i\in A_1} w_i^{(k)} q_i \mathbf{x}_i ,$$

where $w_i^{(k)}$ is the standard jackknife replication weight for unit *i* of the *k*-th replication.

Step 3: Calculate the new jackknife weights for the second phase sample using $\hat{\mathbf{c}}_1^{(k)}$ as control. The form depends on the type of second phase estimator used. Let $\alpha_i^{(k)}$ be the new jackknife replication weight to be determined for unit *i* of the *k*-th replication. The $\alpha_i^{(k)}$ are chosen to minimize

$$\sum_{\substack{i\in A_2 \\ i\neq k}} \left(\alpha_i^{(k)} - w_i^{(k)}\right)^2 \left(w_i^{(k)}\right)^{-1} q_i , \qquad (15)$$

subject to

$$\sum_{i\in A_2} \alpha_i^{(k)} q_i \mathbf{x}_i = \hat{\mathbf{c}}_1^{(k)} , \qquad (16)$$

where $\hat{\mathbf{c}}_1^{(k)}$ is calculated from Step 2, and $w_i^{(k)}$ are the first phase jackknife weights.

The minimizations of Equation (15) subject to the Equation (16) gives the weights

$$\alpha_i^{(k)} = w_i^{(k)} \sum_{g=1}^{G} x_{ig} \left(\frac{\sum_{j\in A_1} w_j^{(k)} q_j x_{jg}}{\sum_{j\in A_2} w_j^{(k)} q_j x_{jg}} \right) . \qquad (17)$$

The jackknife replicate for $\hat{Y}_2$ of the *k*-th replication using weights Equation (17) is

$$\hat{Y}_2^{(k)} = \sum_{i\in A_2} \alpha_i^{(k)} y_i . \qquad (18)$$

The $\hat{Y}_2^{(k)}$ of Equation (18) is algebraically equivalent to the Rao-Shao replicate value for the REE if $q_i = 1$. The replicates for the DEE are constructed with $q_i = w_i^{-1}$.

The procedure can be extended for three-phase sampling. Let a three phase estimator can be written of the form

$$\hat{Y}_3 = \hat{\mathbf{c}}_2' \hat{\mathbf{b}}_3 := \sum_{i\in A_3} \lambda_i y_i ,$$

where $\hat{\mathbf{c}}_2$ is the control total of certain characteristics calculated from the second phase sample and A_3 is the set of indices for the third phase sample. The dimension of the vector $\hat{\mathbf{c}}_2$ is equal to the number of the third phase strata. We assume that we can write

$$\hat{\mathbf{c}}_2 = \sum_{i\in A_2} \alpha_i q_{i2} \mathbf{z}_i ,$$

where α_i is the sampling weight of unit *i* in the second phase sample and $\mathbf{z}_i$ is a vector of the indicator functions for the third phase stratum. Then, in addition to the three steps above we need one more step:

Step 4: Calculate the new jackknife weights for the third phase sample using $\hat{\mathbf{c}}_2^{(k)}$ as control. Let $\lambda_i^{(k)}$ is the new jackknife replication weight for unit *i* of the *k*-th replication. The $\lambda_i^{(k)}$ are chosen to minimize

$$\sum_{\substack{i\in A_3\\ i\neq k}} \left(\lambda_i^{(k)} - \alpha_i^{(k)}\right)^2 \left(\alpha_i^{(k)}\right)^{-1} q_{i2},$$

subject to

$$\sum_{i\in A_3} \lambda_i^{(k)} q_{i2}\mathbf{z}_i = \sum_{i\in A_2} \alpha_i^{(k)} q_{i2}\mathbf{z}_i\,, \qquad (19)$$

where $\alpha_i^{(k)}$ is calculated from Step 3.

4. Application to the 2000 US Census

4.1 Introduction

The Census Bureau will conduct the Accuracy and Coverage Evaluation (ACE) survey after the initial phase of the census enumeration. The ACE will rely on Dual System Estimation (DSE) to determine estimates. For details of the DSE, see Wolter (1986). The estimator is

$$D\hat{S}E = (C - II)\left(\frac{\hat{CE}}{\hat{N}_e}\right)\left(\frac{\hat{N}_n + \hat{N}_i}{\hat{M}_n + \left(\frac{\hat{M}_o}{\hat{N}_o}\right)\hat{N}_i}\right),$$

where C is the census count, II is the number of whole-person census imputations, $\hat{N}_e$ is the estimated E-sample total, $\hat{N}_n$ is the estimated P-sample nonmovers, $\hat{N}_i$ is the estimated P-sample inmovers, $\hat{N}_o$ is the estimated P-sample outmovers, $\hat{M}_n$ is the estimated P-sample nonmover matches, and $\hat{M}_o$ is estimated P-sample outmover matches.

Hence, there are seven components to be estimated to construct a DSE. For the sake of simplicity, we only consider the estimation of the total number of matches. To decide the match status for the individuals in a block, a computerized system is used to check if an individual in the P-sample is also located in the corresponding black of the E-sample. The computer operation of searching in the block is called initial housing unit matching operation.

In the 1990 PES, the surrounding blocks as well as the corresponding block were searched for matches. The search in the surrounding blocks is called the extended search. Extended searches were made for all sample blocks in the 1990 PES. In the 2000 ACE, the extended search sample blocks are selected on information obtained from the initial matching. This search of sample blocks is called Targeted Extended Search (TES).

The ACE survey, which will be performed after the census enumeration, has the following three-phase sampling features.

Phase 1: 29,636 Block Clusters (BC) were selected using the Integrated Coverage Measurement (ICM) survey design.

Phase 2: 11,803 BC were selected from the ACE sample using a stratification of the 30,000 BC.

Phase 3: Checking for a match of a person in the surrounding block area is performed on 20 percent of the ACE sample. This is called the TES.

The overall process was performed to gain efficiency because the available information is different for each phase. Phase Two uses the information of the Census List and Independent List for the ACE survey. Phase Three uses information attained in the initial matching of housing units in the P-sample and the E-sample.

4.2 Point Estimation

Let the parameter of interest be the state population total

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}\,,$$

where y_{hi} is the value of the study variable at the third phase for a given poststratum. Without loss of generality, we assume that the first n_h units are selected as the ICM sample.

The ICM sample is an equal probability sample in each stratum. If y were observed for the entire first phase sample, the full sample estimator would be

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi}\,, \qquad (20)$$

where w_{hi} is the original sampling weight. Since the ICM is a stratified random sample, $w_{hi} = n_h^{-1} N_h$.

The original ICM design has four strata. Stratum one contains small blocks clusters (0-2 HU/BC), stratum two contains medium blocks clusters (3-79 HU/BC), stratum three contains large block clusters (80 or over HU/BC), and stratum four contains American Indian Region block clusters.

A portion of the ICM samples are selected for the ACE sample. The design for the ACE sample uses stratification of the ICM sample. Within each block cluster stratum, the original ICM sample is partitioned into four substrata. They are a consistency substratum, a high inconsistency substratum, a low inconsistency substratum, and a minority substratum. The ACE subsampling was done independently within each substratum.

Define a substratum indicator function x_{hgi}, which takes the value one if unit i in stratum h belongs to the g-th substratum and zero otherwise. Note that $\sum_{g=1}^{G} x_{hgi} = 1$ because each unit belongs to one and only one substratum. The x_{hgi} defines a population characteristic, membership in a substratum.

Also, define I_{hgi} to be the sample selection indicator for second phase selection. The subscript g is used because the selection for the second phase sampling is dependent on substratum g. If the entire second phase sample, the complete ACE sample, is observed, the two-phase estimator would be

$$\hat{Y}_2 = \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi} I_{hgi} y_{hi} , \qquad (21)$$

where

$$\alpha_{hgi} = w_{hi} x_{hgi} \frac{\sum_{j=1}^{n_h} w_{hj} x_{hgj}}{\sum_{j=1}^{n_h} w_{hj} x_{hgj} I_{hgj}} .$$

This is a two-phase estimator where the ACE sample is the second phase sample. However, there is a third phase of sampling.

The study variable is written as $y_{hi} = u_{hi} + v_{hi}$, where u_{hi} is the initial value of the study variable for BC i in stratum h in the state after initial matching and v_{hi} is the change in value due to TES. In the third phase of sampling, the ACE sample is divided into three groups, based on the value of the u_{hi} in the ACE sample. These strata are formed at a national level without consideration for the first phase strata and without consideration for the second phase strata. The first group is called the non-TES group, where all the BC in the non-TES group is excluded from the TES operation. The second group is called the TES-sampling group, where a subsample is selected for TES determination. The third group is called the TES-certainty group, where all BC has a TES determination mode. Define a third phase group indicator function s_{hic}, which takes the value one if unit i in stratum h belongs to the c-th group and zero otherwise. The s_{hic} defines a population characteristic, membership in a third phase group.

Define a_{hic} to be the TES sample selection indicator for membership of unit (hi) in the c-th group. Then, a three-phase estimator of the total of y=u+v is

$$\hat{Y}_3 = \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi} I_{hgi} u_{hi} + \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \sum_{c=2}^{3} \alpha_{hgi} I_{hgi} \alpha_c^* v_{hi}, \qquad (22)$$

where

$$\alpha_c^* = s_{hic} \frac{\sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} I_{hgi} s_{hic}}{\sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} I_{hgi} a_{hic} s_{hic}} .$$

This estimator in Equation (22) uses a difference estimator in moving from the third phase to the second phase. This difference estimator was judged to be a simple and natural estimator in this situation. Only for v_{hi} is the third phase sampling variability. The estimator in Equation (22) is not in the class discussed in Section 1, but each of the two parts is a member of the class. The first part, the part in u, is a two-phase estimator of the REE type. The second part, the part in v, is of the DEE type because the final weight of unit (hi) in the third phase sample is the product of the ACE sample weight α_{hgi} and the TES sampling factor α_c^*.

4.3 Variance Estimation

A set of jackknife replication weights were constructed at the ACE sample level for variance estimation. The number of the replications is equal to the number of BC's in the ICM sample.

Under the two-phase sampling setup, we are able to construct the replication weights for a variance estimator. Let $w_{hi}^{(st)}$ be the standard jackknife replication weight for the first phase sample. That is

$$w_{hi}^{(st)} = \begin{cases} 0 & \text{if } (hi) = (st) \\ \dfrac{n_h}{n_h - 1} w_{hi} & \text{if } h = s, i \neq t\,. \\ w_{hi} & \text{if } h \neq s \end{cases} \tag{23}$$

The replicate for the total of y based on the ACE sample, when the first phase unit (st) is deleted, is

$$\hat{T}_{y_2}^{(st)} = \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi}^{(st)} I_{hgi} y_{hi}\,,$$

where

$$\alpha_{hgi}^{(st)} = w_{hi}^{(st)} x_{hgi} \frac{\sum_{j=1}^{n_h} w_{hj}^{(st)} x_{hgj}}{\sum_{j=1}^{n_h} w_{hj}^{(st)} x_{hgj} I_{hgj}}\,. \tag{24}$$

Notice that only the second phase sample is used to construct the replicates. The weights $\alpha_{hgi}^{(st)}$ are

$$\alpha_{hgi}^{(st)} = \begin{cases} 0 & \text{if } (hi) = (st) \\ \dfrac{r_{hg}}{r_{hg}-1} \dfrac{n_{hg}-1}{n_{hg}} \dfrac{n_h}{n_h-1} \alpha_{hgi} & \text{if } h=s, x_{sgt}=1, I_{sgt}=1, i \neq t \\ \dfrac{n_{hg}-1}{n_{hg}} \dfrac{n_h}{n_h-1} \alpha_{hgi} & \text{if } h=s, x_{sgt}=1, I_{sgt}=0, i \neq t\,, \\ \dfrac{n_h}{n_h-1} \alpha_{hgi} & \text{if } h=s, x_{sgt}=0, i \neq t \\ \alpha_{hgi} & \text{if } h \neq \text{s} \end{cases} \tag{25}$$

where n_h is the first phase sample size of stratum h, n_{hg} is the first phase sample size of substratum g in stratum h, r_{hg} is the second phase sample size of substratum g in stratum h.

Them, the jackknife replicate of $\hat{T}_3$ based on the replication weights in Equations (23) is

$$\hat{T}_{y_3}^{(st)} = \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi}^{(st)} I_{hgi} u_{hi} + \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \sum_{c=2}^{3} \alpha_{hgi}^{(st)} I_{hgi} \alpha_c^{*(st)} a_{hic} v_{hi}\,,$$

where

$$\alpha_c^{*(st)} = s_{hic} \sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi}^{(st)} I_{hgi} s_{hic} \alpha_{hgi}^{-1} \times \left[\sum_{h=1}^{H} \sum_{g=1}^{G} \sum_{i=1}^{n_h} \alpha_{hgi}^{(st)} I_{hgi} a_{hic} s_{hic} \alpha_{hgi}^{-1} \right]^{-1}\,.$$

Variances based on these replicates will establish the precision of estimates calculated in the Accuracy and Coverage Evaluation of the 2000 Decennial Census.

5. References

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

Fuller, W.A. (1998). Replication Variance Estimation for Two-Phase Samples. *Statistica Sinica*, **8**, pp. 1153-1164.

Kott, P.S. (1990). Variance Estimation When a First-Phase Area Sample is Re-stratified. *Survey Methodology*, **16**, pp. 99-103.

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

Rao, J.N.K. (1973). On Double Sampling for Stratification and Analytical Surveys. *Biometrika*, **60**, pp. 125-133.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation. *Biometrika*, **79**, pp. 811-822.

Wolter, K. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, **81**, pp. 338-346.

Acknowledgments

The work was performed while the first author was on appointment as a mathematical statistician in the Decennial Statistical Studies Division at the U.S. Census Bureau. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

The first author is grateful to Chantell Adams for repeated typing of the manuscript.