Statistica Sinica **8**(1998), 1153-1164

REPLICATION VARIANCE ESTIMATION FOR TWO-PHASE SAMPLES

Wayne A. Fuller

Iowa State University

Abstract: The estimation of the variance of the regression estimator for a two-phase sample is considered. Given the covariance matrix of the first-phase control variables, a replication variance estimator that uses only the second-phase sample is developed. The procedure has computational advantages for surveys with second-phase samples that are small relative to the first-phase sample.

Key words and phrases: Balanced repeated replication, double sampling, regression estimator, sample surveys.

1. Introduction

Two-phase samples are samples in which a vector of characteristics, denoted by $\boldsymbol{X}_t$ where t denotes an individual element, is observed on a relatively large sample. An extended vector, denoted by $(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ is observed on a subsample of the large sample. We denote the first sample by $s_{(1)}$ and the second sample by $s_{(2)}$. Estimation for two-phase samples is discussed in texts such as Cochran (1977), Särndal, Swensson and Wretman (1992), and Wolter (1985). Different procedures can be used to combine the information from the first and second samples to obtain estimators of functions of $(\boldsymbol{X}, \boldsymbol{Y})$. We will concentrate on the regression method.

Variance estimation for two-phase samples based on standard Taylor arguments is described in the references previously cited. A jackknife procedure has been suggested for the ratio estimator by Rao and Sitter (1995) and for the regression estimator by Sitter (1997). These procedures involve creating jackknife replicates using the entire first-phase sample. Särndal and Swensson (1987) discuss variance estimation for regression estimation in the general two-phase situation. Kott (1990, 1995), discussed variance estimation in a situation where the stratified design for the second-phase differs from the stratified design for the first-phase. He suggests a jackknife estimator of variance for a particular two-phase estimator. Breidt and Fuller (1993) suggested a replication variance estimator for multiphase samples.The Breidt and Fuller procedure is particularly applicable if the first-phase and second-phase primary sampling units are the

same and if there are at least two second-phase primary sampling units in each first-phase stratum.

In the classical two-phase situation the investigator specifies the selection probabilities for both sample selections. The methodology of two-phase samples is also used for situations in which the second-phase selection probabilities are not under the control of the investigator. One example is the case of nonresponse. See Särndal and Swensson (1987), Fuller, Loughin and Baker (1994), and An, Breidt and Fuller (1994). The two-phase model has also been used in biometry and epidemiology. Unfortunately, the term two-stage sampling has been used by authors in those areas. See Flander and Greenland (1991) and Zhao and Lipsitz (1992).

2. Variance Estimation

We study the situation in which the selection probabilities are known. We shall investigate estimation of the mean of $\boldsymbol{Y}$ using the regression estimator. Let

$\hat{\boldsymbol{\mu}}_{X(1)}$= estimator of the population mean of $\boldsymbol{X}$ constructed from $s_{(1)}$.
$\hat{\boldsymbol{\mu}}_{X(2)}$= estimator of the mean of $\boldsymbol{X}$ constructed from $s_{(2)}$. This can also be considered an estimator of $\hat{\boldsymbol{\mu}}_{X(1)}$.
$\hat{\boldsymbol{\mu}}_{Y(2)}$= estimator of the mean of $\boldsymbol{Y}$ constructed from $s_{(2)}$.
$(\boldsymbol{\mu}_{X(F)}, \boldsymbol{\mu}_{Y(F)})$= vector of finite population means.

We assume a sequence of samples and finite populations such as that described in Fuller (1975). We let the sequence be indexed by n and assume that the size of the finite populations, denoted by N_n, increases as the sample size n increases such that the limit of $N_n^{-1}n$ is a finite fraction, perhaps zero. The sample moments of the finite population are assumed to converge to finite constants such that the difference between the moments for the nth finite population and the limiting moments is $O_p(N_n^{-1/2})$. We assume

$$E\{\hat{\boldsymbol{\mu}}_{X(1)n},\ \hat{\boldsymbol{\mu}}_{X(2)n}, \hat{\boldsymbol{\mu}}_{Y(2)n}\} = (\boldsymbol{\mu}_{X(F)},\ \boldsymbol{\mu}_{X(F)}, \boldsymbol{\mu}_{Y(F)}), \tag{2.1}$$

$$V\{(\hat{\boldsymbol{\beta}}_{(2)n} - \boldsymbol{\beta}_{(F)n})',\ \hat{\boldsymbol{\mu}}_{X(1)n},\ \hat{\boldsymbol{\mu}}_{X(2)n},\ \hat{\boldsymbol{\mu}}_{Y(2)n}\} = O(n^{-1}), \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}_{(2)n}$ is an estimator of the vector of coefficients in the regression of Y on $\boldsymbol{X}$ computed from sample two and $\boldsymbol{\beta}_{(F)n}$ is the finite population analog of $\hat{\boldsymbol{\beta}}_{(2)n}$. For simplicity, we consider a single Y and omit the subscript n in the remainder of the discussion. All variance expressions extend immediately to covariance matrices for a vector of regression estimators based on the common vector $\boldsymbol{X}$.

Our estimator of $\mu_{Y(F)}$ is

$$\hat{\mu}_{Y(F)} = \hat{\mu}_{Y(2)} + (\hat{\boldsymbol{\mu}}_{X(1)} - \hat{\boldsymbol{\mu}}_{X(2)})\hat{\boldsymbol{\beta}}_{(2)}\,, \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}_{(2)}$ is the coefficient vector computed from $s_{(2)}$. If $(Y_j, \boldsymbol{X}_j), j=1,2,...,n_2$, are the observations on the second-phase sample, an estimator of $\boldsymbol{\beta}_{(2)}$ is

$$\hat{\boldsymbol{\beta}}_{(2)} = \Big(\sum_{t=1}^{n_2} \pi_t^{-1} \ \boldsymbol{X}_t' \ \boldsymbol{X}_t\Big)^{-1} \sum_{t=1}^{n_2} \pi_t^{-1} \ \boldsymbol{X}_t' Y_t \,,$$

where π_t^{-1} is the sampling weight. Usually, the sampling weight is the inverse of the selection probability, where the selection probability is the first-phase selection probability multiplied by the conditional second-phase selection probability given the first-phase sample. We assume that the estimator $\hat{\boldsymbol{\beta}}_{(2)}$ is such that the estimator of the mean given in (2.3) can also be written

$$\hat{\mu}_{Y(F)} = \sum_{t=1}^{n_2} w_t Y_t, \tag{2.4}$$

where

$$\sum_{t=1}^{n_2} w_t \boldsymbol{X}_t = \hat{\boldsymbol{\mu}}_{X(1)}, \tag{2.5}$$

n_2 is the number of units in the second-phase sample, and w_t is the regression weight. The representation (2.4) will hold for a large class of regression estimators, $\hat{\boldsymbol{\beta}}_{(2)}$. A common procedure for cluster samples is to estimate $\boldsymbol{\beta}_{(F)}$ using the observation unit regression of Y on $\boldsymbol{X}$, instead of the cluster total regression. While the representation (2.4) holds for such a procedure, the estimator may not be fully efficient. If there are a large number of sampling units and the dimension of $\boldsymbol{X}$ is small, so that the large sample approximations are appropriate, the variance of the estimator is minimized by computing $\boldsymbol{\beta}$ as the regression based on the totals of the sampling units. See Rao (1994).

If the error in the estimators of moments is $O_p(n^{-1/2})$ and if $\hat{\boldsymbol{\beta}}_{(2)}$ is a function of estimated moments, then $\hat{\boldsymbol{\beta}}_{(2)} - \boldsymbol{\beta}_{(F)} = O_p(n^{-1/2})$, and

$$\hat{\mu}_{Y(F)} - \mu_{Y(F)} = \hat{\mu}_{e(2)} + (\hat{\boldsymbol{\mu}}_{X(1)} - \boldsymbol{\mu}_{X(F)})\boldsymbol{\beta}_{(F)} + O_p(n^{-1}), \tag{2.6}$$

where $e_t = Y_t - \mu_{Y(F)} - (\boldsymbol{X}_t - \boldsymbol{\mu}_{X(F)})\boldsymbol{\beta}_{(F)}$ and

$$\hat{\mu}_{e(2)} = \hat{\mu}_{Y(2)} - \mu_{Y(F)} + (\boldsymbol{\mu}_{X(F)} - \hat{\boldsymbol{\mu}}_{X(2)})\boldsymbol{\beta}_{(F)} \,. \tag{2.7}$$

We call $\hat{\mu}_{e(2)} + (\hat{\boldsymbol{\mu}}_{X(1)} - \mu_{X(F)})\, \boldsymbol{\beta}_{(F)}$ the leading term of (2.6). See Fuller (1975). In most two-phase procedures, the second-phase estimator of the mean of $\boldsymbol{X}$ is constructed to be unbiased, or consistent, for the first phase mean. We assume

$$E\{\hat{\boldsymbol{\mu}}_{X(2)} - \hat{\boldsymbol{\mu}}_{X(1)} | s_{(1)}\} = \mathbf{0}, \tag{2.8}$$

or the weaker condition

$$E\{\hat{\boldsymbol{\mu}}_{X(2)} - \hat{\boldsymbol{\mu}}_{X(1)}|s_{(1)}\} = O_p(n^{-1}).$$

Under assumption (2.8),

$$\begin{aligned} C\{\hat{\mu}_{e(2)},\ \hat{\boldsymbol{\mu}}_{X(1)}\} &= E\{C[\hat{\mu}_{e(2)} - \hat{\mu}_{e(1)},\ \hat{\boldsymbol{\mu}}_{X(1)}|s_{(1)}]\} \\ &\quad + C\{E[\hat{\mu}_{e(2)} - \hat{\mu}_{e(1)}|s_{(1)}],\ E[\hat{\boldsymbol{\mu}}_{X(1)}|s_{(1)}]\} + C\{\hat{\mu}_{e(1)},\ \hat{\boldsymbol{\mu}}_{X(1)}\} \\ &= C\{\hat{\mu}_{e(1)},\ \hat{\boldsymbol{\mu}}_{X(1)}\}. \end{aligned} \tag{2.9}$$

If

$$C\{\hat{\mu}_{e(1)},\ \hat{\boldsymbol{\mu}}_{X(1)}\} = \mathbf{0}\,, \tag{2.10}$$

and (2.8) holds, then

$$V\{\hat{\mu}_{Y(F)} - \mu_{Y(F)}\} \doteq V\{\hat{\mu}_{e(2)}\} + \boldsymbol{\beta}'_{(F)}\boldsymbol{V}_{XX11}\boldsymbol{\beta}_{(F)}, \tag{2.11}$$

where $\boldsymbol{V}_{XX11}$ denotes the covariance matrix of $\hat{\boldsymbol{\mu}}_{X(1)} - \boldsymbol{\mu}_{X(F)}$.The covariance between $\hat{\mu}_{e(1)}$ and $\hat{\boldsymbol{\mu}}_{X(1)}$ will be zero if the regression equation is calculated using the first-phase primary sampling unit totals as the observations in the second-phase regression. Alternatively, for example, if the first-phase sample is a cluster sample and if the second-phase regression is computed using the elements as observations, then it is possible for $\hat{\mu}_{e(2)}$ and $\hat{\mu}_{X(1)}$ to be correlated.

Conditioning on $\hat{\boldsymbol{\mu}}_{X(1)}$, not on the entire first-phase sample, the variance of the leading term of (2.6) is

$$\begin{aligned} V\{\hat{\mu}_{Y(F)}\} \doteq\ & E\{V[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}]\} + V\{E[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}]\} \\ & + 2C\{E[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}],\ \hat{\boldsymbol{\mu}}_{X(1)}\boldsymbol{\beta}_{(F)}\} + V\{\hat{\boldsymbol{\mu}}_{X(1)}\boldsymbol{\beta}_{(F)}\}. \end{aligned} \tag{2.12}$$

If $V\{E[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}]\} = o\left(n^{-1}\right)$ then

$$V\{\hat{\mu}_{Y(F)}\} \doteq E\{V[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}]\} + \boldsymbol{\beta}'_{(F)}\boldsymbol{V}_{XX11}\boldsymbol{\beta}_{(F)}, \tag{2.13}$$

where terms in the variance expression (2.12) of smaller order than $O\left(n^{-1}\right)$ are omitted to obtain the approximation. For stratification at the second-phase, $\hat{\boldsymbol{\mu}}_{X(1)}$ is essentially the vector of estimated fractions in the second-phase strata. In this case, (2.12) is satisfied because the conditional expectation in (2.12) is zero. See Section 3.

Thus, conceptually, it is easy to use (2.11) or (2.13) to define an estimator of the variance of $\hat{\mu}_{Y(F)}$. One computes the regression of Y on $\boldsymbol{X}$ in sample two to obtain the estimator $\hat{\boldsymbol{\beta}}_{(2)}$. Then one uses any standard estimating procedure to estimate the variance of the mean of $Y_t - \boldsymbol{X}_t\hat{\boldsymbol{\beta}}_{(2)}$ as an estimator of the population mean. Given an estimator of $\boldsymbol{V}_{XX11}$, denoted by $\hat{\boldsymbol{V}}_{XX11}$, one

adds $\hat{\boldsymbol{\beta}}'_{(2)}\hat{\boldsymbol{V}}_{XX11}\hat{\boldsymbol{\beta}}_{(2)}$, to the estimated variance of $\hat{\mu}_{e(2)}$. Unfortunately, this is a cumbersome computational procedure when there are a large number of Y's because the regression coefficient vector must be computed for each Y.

Our objective is to design a variance estimation scheme that is computationally less burdensome. Assume that replicate weights have been constructed for the second-phase sample to estimate the variance of $\hat{\mu}_{e(2)}$. These can be balanced repeated replication weights or jackknife weights. Depending upon the situation, one may construct replicates to estimate the first term of (2.11) or one may construct replicates to estimate $V\{\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}\}$. The second procedure can be used when the estimated conditional variance is also an estimator of the unconditional variance. Let there be a set of r replicates such that the estimator for the ith replicate is $\hat{\mu}_{Y(2)i} = \sum_{t=1}^{n_2} w_{it}Y_t$, where Y_t is the observation on element t for any Y characteristic and w_{it} is the weight of replicate i for element t. Assume

$$E\Big\{\sum_{i=1}^{r}(\hat{\mu}_{e(2)i} - \hat{\mu}_{e(2)})^2\Big\} = cV\{\hat{\mu}_{e(2)}\} + O_p(n^{-3/2}), \tag{2.14}$$

for some fixed c.

We now obtain a new set of weights that can be used to estimate the variance of the two-phase estimator. We assume the number of replicates, denoted by r, is greater than or equal to k, the dimension of $\boldsymbol{X}$. Let $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_r$ be a set of k-dimensional row vectors with the property that

$$\sum_{i=1}^{r}\boldsymbol{\delta}'_i\boldsymbol{\delta}_i = c\hat{\boldsymbol{V}}_{XX11}, \tag{2.15}$$

where c is the constant defined in (2.14). It is always possible to construct $\boldsymbol{\delta}_i, i = 1, \ldots, r$, satisfying (2.15) if $r \geq k$. For example, if $\boldsymbol{q}'_i,\ i = 1, \ldots, k$, are the characteristic vectors of $\hat{\boldsymbol{V}}_{XX11}$ and λ_i are the corresponding roots, then $\boldsymbol{\delta}_i = (c\lambda_i)^{1/2}\boldsymbol{q}_i,\ i = 1, \ldots, k$, and $\boldsymbol{\delta}_i = \mathbf{0},\ i = k+1,\ k+2,\ \ldots, r$ satisfy (2.15).

Define a new set of replication weights, denoted by a_{it}, and the associated estimators

$$\tilde{\mu}_{Yi} = \sum_{t=1}^{n_2} a_{it}Y_t = \hat{\mu}_{Y(2)i} + (\hat{\boldsymbol{\mu}}_{X(1)} + \boldsymbol{\delta}_i - \hat{\boldsymbol{\mu}}_{X(2)i})\hat{\boldsymbol{\beta}}_{(2)i}, \tag{2.16}$$

where $\hat{\boldsymbol{\beta}}_{(2)i}$ is the regression coefficient from the regression of Y on $\boldsymbol{X}$ in the ith replicate of sample two using the weights w_{it} as initial weights, and the a_{it} satisfy

$$\sum_{t=1}^{n_2} a_{it}\boldsymbol{X}_t = \hat{\boldsymbol{\mu}}_{X(1)} + \boldsymbol{\delta}_i, \quad i = 1, \ldots, r. \tag{2.17}$$

Then

$$\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)} = \hat{\mu}_{e(2)i} - \hat{\mu}_{e(2)} + \boldsymbol{\delta}_i \boldsymbol{\beta}_{(F)} + O_p(n^{-1}). \tag{2.18}$$

If the $\boldsymbol{\delta}_i$ are randomly assigned to the replicates, and if the $O_p(n^{-1})$ term in (2.18) is ignored,

$$E\Big\{ \sum_{i=1}^{r} (\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)})^2 \Big\} \doteq cV\{\hat{\mu}_{e(2)}\} + c\boldsymbol{\beta}'_{(F)} \boldsymbol{V}_{XX11} \boldsymbol{\beta}_{(F)}. \tag{2.19}$$

Therefore, a replication estimator of the variance of the two-phase estimator is

$$\hat{V}\{\hat{\mu}_Y\} = c^{-1} \sum_{i=1}^{r} \left(\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)} \right)^2, \tag{2.20}$$

where $\tilde{\mu}_{Yi}$ is defined in (2.16).

The replication estimator has two desirable features. Given the weights a_{it}, only the second-phase sample is used in the calculations for $\hat{Y}$ and the variance of $\hat{Y}$. This is a particular advantage for samples in which the first phase is large relative to the second phase. Second, the estimated variance for the characteristic $\boldsymbol{X}$ is the variance estimated from the first-phase sample.

We now examine the selection of $\boldsymbol{\delta}_i$. Assume a two-per-stratum design with normally distributed observations. If jackknife replicates are created by deleting one unit in a single stratum and doubling the weight of the other, the jackknife difference for the y-variable for stratum i is $w_i (y_{i1} - y_{i2})$ when the second unit in stratum i is deleted and the weight for stratum i is w_i. Thus, the original jackknife differences before the regression adjustment are independent in this case. The jackknife difference for the regression estimator can be written

$$\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)} = \varepsilon_i + \boldsymbol{\delta}_i \hat{\boldsymbol{\beta}}_{(2)i} = \varepsilon_i + \boldsymbol{\delta}_i \hat{\boldsymbol{\beta}}_{(2)} + O_p(n^{-3/2}), \tag{2.21}$$

where $\varepsilon_i = O_p(n^{-1})$. Now

$$\sum_{i=1}^{r} (\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)})^2 = \sum_{i=1}^{r} (\varepsilon_i + \boldsymbol{\delta}_i \boldsymbol{\beta}_{(F)})^2 + O_p(n^{-3/2}), \tag{2.22}$$

where $r = 0.5n_2$ is the number of strata. If the ε_i are normally distributed, $V\{\varepsilon_i^2\} = 2\sigma_{\varepsilon_i}^4$. Then, under the assumptions that $\boldsymbol{\delta}_i$ and ε_i are independent, and that $\hat{\boldsymbol{V}}_{XX11}$ is independent of the ε_i,

$$V\Big\{ \sum_{i=1}^{r} (\varepsilon_i + \boldsymbol{\delta}_i \boldsymbol{\beta}_{(F)})^2 \Big\} = 2 \sum_{i=1}^{r} \sigma_{\varepsilon_i}^4 + 4 \sum_{i=1}^{r} \sigma_{\varepsilon_i}^2 \boldsymbol{\beta}'_{(F)} \boldsymbol{V}_{XX11} \boldsymbol{\beta}_{(F)} + V\{\boldsymbol{\beta}'_{(F)} \hat{\boldsymbol{V}}_{XX11} \boldsymbol{\beta}_{(F)}\}. \tag{2.23}$$

This computation suggests that the form of the $\boldsymbol{\delta}_i$ is not overly important, given that they satisfy (2.15) and are randomly assigned to replicates.

The second term in (2.23) can be eliminated by creating a set of replicates balanced on the $\boldsymbol{\delta}_i$. To illustrate, assume r initial replicates, where $r \geq k$, and assume (2.14) and (2.15). Define $2r + k$ replicates by

$$\begin{aligned}
\tilde{\mu}_{Yi1} &= \hat{\mu}_{Y(2)i} + (\hat{\boldsymbol{\mu}}_{X(1)} + \boldsymbol{\delta}_i - \hat{\boldsymbol{\mu}}_{X(2)i})\hat{\boldsymbol{\beta}}_{(2)i},\ i = 1, \ldots, k \\
\tilde{\mu}_{Yi2} &= \hat{\mu}_{Y(2)i} + (\hat{\boldsymbol{\mu}}_{X(1)} - \boldsymbol{\delta}_i - \hat{\boldsymbol{\mu}}_{X(2)i})\hat{\boldsymbol{\beta}}_{(2)i},\ i = 1, \ldots, k \\
\tilde{\mu}_{Yi} &= \hat{\mu}_{Y(2)i} + (\hat{\boldsymbol{\mu}}_{X(1)} - \hat{\boldsymbol{\mu}}_{X(2)i})\hat{\boldsymbol{\beta}}_{(2)i}, \qquad i = k+1, \ldots, r.
\end{aligned} \tag{2.24}$$

The estimator of the variance is

$$\begin{aligned}
\hat{V}\{\hat{\mu}_{Y(F)}\} &= c^{-1}\Big\{0.5\sum_{i=1}^{k}[(\tilde{\mu}_{Yi1} - \hat{\mu}_{Y(F)})^2 + (\tilde{\mu}_{Yi2} - \hat{\mu}_{Y(F)})^2] + \sum_{i=k+1}^{r}(\tilde{\mu}_{Yi} - \hat{\mu}_{Y(F)})^2\Big\} \\
&= c^{-1}\sum_{i=1}^{r}(\tilde{\mu}_{ei} - \hat{\mu}_{e(F)})^2 + \hat{\boldsymbol{\beta}}'_{(F)}\hat{\boldsymbol{V}}_{XX11}\hat{\boldsymbol{\beta}}_{(F)} + O_p(n^{-3/2})
\end{aligned} \tag{2.25}$$

and the efficiency of the variance estimator depends only on the efficiency of the second-phase replication procedure.

3. Second-Phase Replication Procedures

In constructing the variance estimator, we assumed it is possible to create replicates for the computation of an estimator of the unconditional variance for the estimated population mean of e_t from the second-phase sample. We give some situations in which unbiased or consistent variance estimators can be constructed and suggest approximations for other designs.

Simple random sampling at both phases. In this case, the second-phase sample is a simple random sample from the entire population. Hence, any replicate procedure appropriate for simple random sampling can be used to construct replicates. A stratified first phase sample with a simple random second-phase sample in each first phase stratum is an immediate extension of two phases of simple random sampling.

Two-phase with second-phase stratified. Consider a two-phase sample with a simple random sample first-phase and a second-phase that is a stratified sample of the first-phase sample. Assume that the estimator is the stratified estimator using estimated population sizes from the first-phase. The second-phase stratified estimator is equivalent to the regression estimator (2.3) in which indicator variables for the second-phase strata are the elements of the $\boldsymbol{X}$-vector. Given such X-variables, the mean of the e-variables is zero for every second-phase stratum, and $E\{\hat{\mu}_{e(2)} | (n_{11}, \ldots, n_{1H})\} = 0$, where n_{1j} is the number of first-phase units that fall in second-phase stratum j, $\hat{\boldsymbol{\mu}}_{X(1)} = n^{-1}(n_{11}, \ldots, n_{1H})$, and H is the number of second-phase strata. It follows that the variance of $\hat{\mu}_{e(2)}$ is the expected value of the conditional variance given $\hat{\boldsymbol{\mu}}_{X(1)}$. Therefore, replicates

constructed to give an unbiased estimator of the conditional variance also give an unbiased estimator of the unconditional variance. Observe that the first-phase sample need not be a simple random sample. The requirement is that the conditional expected value of $\hat{\mu}_{e(2)}$ given $\hat{\boldsymbol{\mu}}_{X(1)}$ be zero.

Poisson Sampling. In some situations the observations on the first-phase are used to determine the probabilities of selecting units at the second-phase, stratification being only one example. To investigate estimation for such situations, let a Poisson first-phase sample be selected from a population of N elements with probabilities p_{1t}. From that Poisson sample, select a second-phase Poisson sample with conditional probabilities p_{2t}. Then the second-phase sample is a Poisson sample from the original population with selection probabilities $\pi_t = p_{1t}p_{2t}$, where π_t is the inverse of the sampling weight. It is permissible for p_{2t} to be a fixed function of the value of an X-vector, perhaps only observed on the elements selected at the first-phase. Let the population mean be estimated with

$$\hat{\mu} = \Big(\sum_{s(2)} \pi_t^{-1}\Big)^{-1}\Big(\sum_{s(2)} \pi_t^{-1} Y_t\Big), \tag{3.1}$$

where $\sum_{s(2)}$ denotes the summation over elements in the second-phase sample. Let $n_{p(2)}$ be the second-phase expected sample size and assume

$$\begin{aligned} E\{(\hat{\mu}-\mu)^2\} &= O(n_{p(2)}^{-1}), \\ E\left\{(Y_t-\mu)\left(\hat{\mu}-\mu\right)\right\} &= O(n_{p(2)}^{-1}) \qquad \text{for all } t. \end{aligned}$$

Then, the usual Taylor approximation for the error in $\hat{\mu}$ of (3.1) is

$$\hat{\mu}-\mu \doteq \Big[E\{\sum_{s(2)} \pi_t^{-1}\}\Big]^{-1} \sum_{s(2)} \pi_t^{-1}\left(Y_t-\mu\right), \tag{3.2}$$

and an estimator of the variance is

$$\hat{V}_T\left(\hat{\mu}\right) = \Big(\sum_{s(2)} \pi_t^{-1}\Big)^{-2} \sum_{s(2)} \Big[\pi_t^{-1}\left(Y_t-\hat{\mu}\right)\Big]^2 \left(1-\pi_t\right). \tag{3.3}$$

Observe that variability due to the random sample size in Poisson sampling has little impact on the variance of the mean and Poisson sampling furnishes a good approximation for nonreplacement simple random sampling. See Hajek (1960). If the second-phase sampling rates are fixed in advance, then Poisson sampling furnishes a good approximation to two-phase schemes in which the second-phase rate is a function of the first-phase X-values.

The approximation is further improved if X-variables are used to construct a regression estimator. Assume that

$$\pi_t^{-1} = (1,\ \boldsymbol{X}_t)\,\boldsymbol{\alpha}, \tag{3.4}$$

where $\boldsymbol{X}_t$ is the vector used to construct estimator (2.3) and $\boldsymbol{\alpha}$ is a fixed vector. Then $E\{\pi_t^{-1}e_t\} = 0$, where e_t is defined in (2.6). If, furthermore, $E\{e_t|\pi_t\} = 0$, then $E\{\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{X(1)}\} = 0$ and the unconditional variance is the expectation of the conditional variance. The expectation $E\{e_t|\pi_t\} = 0$ when the π_t are constant within categories defined by $\boldsymbol{X}$. Therefore, (3.3) is a useful approximation to the unconditional variance of e_t in an extended class of designs in which second-phase selection probabilities are functions of the first-phase X-values. Jackknife or half sample replicates can be used to construct estimators of the variance.

Table 1. Observations in second-phase sample.

	Variable								Weights	
Category	Y	Z	C_1	C_2	C_3	C_4	C_5	C_6	w_t	a_{1t}
1	6.122	5.024	1	0	0	0	0	0	0.098	0.000
1	4.614	3.577	1	0	0	0	0	0	0.136	0.217
2	6.685	5.974	0	1	0	0	0	0	0.084	0.087
2	4.806	5.335	0	1	0	0	0	0	0.096	0.085
3	6.072	7.082	0	0	1	0	0	0	0.043	0.069
3	5.599	3.979	0	0	1	0	0	0	0.090	0.062
4	5.670	4.050	0	0	0	1	0	0	0.112	0.075
4	8.297	7.633	0	0	0	1	0	0	0.048	0.085
5	8.015	8.020	0	0	0	0	1	0	0.076	0.073
5	8.990	8.817	0	0	0	0	1	0	0.064	0.074
6	7.099	7.954	0	0	0	0	0	1	0.043	0.044
6	8.131	8.721	0	0	0	0	0	1	0.037	0.045
7	11.867	10.488	0	0	0	0	0	0	0.040	0.041
7	12.242	11.332	0	0	0	0	0	0	0.033	0.042

4. Illustration

As an example of the variance computations, we consider a two-phase sample in which the first-phase is a simple random sample. In the first-phase sample, the observations are placed in one of seven categories and observations are made on a characteristic, denoted by Z. The second-phase sample is a stratified sample of the first-phase with two observations in each of the seven categories. Table 1 contains the fourteen observations in the second-phase sample. The phase one mean of 150 observations is

$$\hat{\boldsymbol{\mu}}_{X(1)} = (\ 6.1084,\ 0.2333,\ 0.1800,\ 0.1333,\ 0.1600,\ 0.1400,\ 0.0800)\ ,$$

where the first entry is the mean of Z and the last six entries in the vector are the fractions in the first six categories.

Let $\boldsymbol{X} = (Z,\ \boldsymbol{C})$, where $\boldsymbol{C}$ is the vector of dummy variables for the categories defined in Table 1 and Z is the continuous regression variable. The covariance between e and $\boldsymbol{X}$ is zero by the property of regression residuals. Thus, either the unconditional or conditional form of the variance estimator can be used. The variance of the leading term in (2.6) can be written

$$\begin{aligned} &V\left\{\hat{\mu}_{e(2)} + (\hat{\boldsymbol{\mu}}_{X(1)} - \boldsymbol{\mu}_{X(F)})\boldsymbol{\beta}_{(F)}\right\} \\ &= E\left\{V[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{C(1)}]\right\} + V\left\{E[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{C(1)}]\right\} + V\{\hat{\boldsymbol{\mu}}_{X(1)}\boldsymbol{\beta}\}. \end{aligned}$$

The conditional expectation, $E[\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{C(1)}]$, is zero because $\boldsymbol{C}$ is a vector of dummy variables for categories and the mean of e is zero for each category. Therefore, the variance of $\hat{\mu}_{e(2)}$ can be estimated by estimating the variance conditional on $\hat{\boldsymbol{\mu}}_{C(1)}$. This is equivalent to estimating the variance for the second-phase sample, treating the sample as a stratified sample with the categories as strata. We assume the first-phase sample is a small fraction of the population and, hence, ignore the finite population correction. Because the conditioning for the second-phase sample is on $\hat{\boldsymbol{\mu}}_{C(1)}$, not on the entire sample, no finite population correction is required for the conditional variance. If a finite population correction were relevant, the estimated rate of $2\hat{N}_h^{-1}$ could be used for the hth second-phase stratum, where $\hat{N}_h$ is the first-phase estimate of the population number in stratum h.

The estimated mean of Y as defined by (2.4) is 6.718, where the vector of weights is given as w_t in the next-to-last column of Table 1. The regression estimator was computed using the inverse of the category sampling rates as initial weights. If we estimate the two terms in (2.13), we have

$$\begin{aligned} \hat{V}\{\hat{\mu}_{Y(F)}\} &= \hat{V}\{\hat{\mu}_{e(2)}|\hat{\boldsymbol{\mu}}_{C(1)}\} + \hat{\boldsymbol{\beta}}'_{(F)}\hat{\boldsymbol{V}}_{XX11}\hat{\boldsymbol{\beta}}_{(F)} \\ &= 0.0353 + 0.0330 = 0.0683, \end{aligned}$$

where the first term is the sum of squares of the weighted regression residuals divided by six. The first term can be considered to be an estimator conditional on $\hat{\boldsymbol{\mu}}_{C(1)}$ or on $\hat{\boldsymbol{\mu}}_{X(1)}$.

Table 2 contains seven vectors $\boldsymbol{\delta}_i$ such that $\Sigma\boldsymbol{\delta}'_i\boldsymbol{\delta}_i = \hat{\boldsymbol{V}}_{XX11}$. The vectors are $\lambda_i^{0.5}\boldsymbol{q}_i$, where $\boldsymbol{q}'_i$ are the characteristic vectors of $\hat{\boldsymbol{V}}_{XX11}$ and the λ_i are the characteristic roots. The vector of the diagonal elements of $\hat{\boldsymbol{V}}_{XX11}$ is

$$0.01\,(5.018,\ 0.120,\ 0.099,\ 0.078,\ 0.090,\ 0.081,\ 0.049)\,.$$

Seven replicates were created, where the ith replicate was formed by deleting the first element in the ith stratum. Seven sets of replicate weights were constructed

using $\hat{\boldsymbol{\mu}}_{X(1)} + \boldsymbol{\delta}_i$, $i = 1, 2, \ldots, 7$, in (2.16). The weights associated with replicate one of Table 2 are given in the last column of Table 1. These weights applied to Z yield $6.3324 = 6.1084 + 0.2240$, where 0.2240 is the first entry in Table 2 and 6.1084 is the first phase mean. The weights applied to C_1 give $0.217 = 0.233 - 0.016$, where -0.016 is the second element of the first row of Table 2.

Table 2. Vectors for the construction of replicates.

Replicate	$\boldsymbol{\delta}$-Vectors						
1	0.2240	-0.0160	-0.0073	-0.0019	-0.0010	0.0070	0.0087
2	-0.0011	-0.0279	0.0228	0.0035	0.0058	-0.0011	-0.0018
3	-0.0007	-0.0072	-0.0167	0.0038	0.0272	-0.0038	-0.0020
4	-0.0007	-0.0068	-0.0095	0.0259	-0.0108	0.0042	-0.0018
5	-0.0011	-0.0050	-0.0043	-0.0076	0.0003	0.0267	-0.0070
6	-0.0014	-0.0055	-0.0042	-0.0029	-0.0020	0.0024	0.0188
7	0.0002	0.0033	0.0031	0.0030	0.0029	0.0025	0.0021

The estimated variance of $\hat{\mu}_Y$ using seven replicates is 0.0590. If an additional seven replicates are created using $-\boldsymbol{\delta}_i$ in place of $\boldsymbol{\delta}_i$, the estimated variance calculated from the 14 replicates is 0.0701. The jackknife variance estimate is slightly larger than the Taylor estimate of 0.0683. This is the usual ordering of the size of the two estimation procedures.

Acknowledgements

This research was supported in part by Cooperative Agreements 68-3A75-43 and 68-3A75-4-86 between the United States Department of Agriculture, Natural Resources Conservation Service and Iowa State University.

References

An, A. B., Breidt, F. J. and Fuller, W. A. (1994). Regression weighting methods for SIPP data. *Proc. ASA Section on Survey Research Methods*, 434-439.

Breidt, F. J. and Fuller, W. A. (1993). Regression weighting for multiphase samples. *Sankhyā Ser. B* **55**, 297-309.

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

Flander, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med.* **10**, 739-747.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya C* **37**, 117-132.

Fuller, W. A., Loughin, M. M. and Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 National Food Consumption Survey. *Survey Methodology* **20**, 75-85.

Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Inst. Hung. Acad. Sci.* **5**, 361-374.

Kott, P. S. (1990). Variance estimation when a first-phase area sample is restrat-ified. *Survey Methodology* **16**, 99-103.

Kott, P. S. (1995). Can the jackknife be used with a two-phase sample? *Proceedings of the Survey Research Methods Section*, Statist. Soc. Canada, 107-110.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Statist.* **10**, 153-165.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**, 453-460.

Särndal, C. E. and Swensson, B. (1987). A general view of estimation for two-phases of selection with applications to two-phase sampling and non-response. *Int'l. Statist. Review* **55**, 279-294.

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

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *J. Amer. Statist. Assoc.* **92**, 780-787.

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

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statist. Med.* **11**, 769-782.

Statistical Laboratory and Department of Statistics, Snedecor Hall, Iowa State University, Ames, IA 50011-1210, U.S.A.

E-mail: waf@iastate.edu

(Received October 1996; accepted June 1998)