Chapter 11

Two-phase sampling

11.1 Introduction

Two-phase sampling design is a sampling design where the sample selection is performed in two phases, where in the first phase the auxiliary variable $\mathbf{x}$ is observed and in the second phase the study variable y is observed. Two-phase sampling is particularly useful when the cost for observing $\mathbf{x}$ is relatively cheap compared with the cost for observing y. To formalize, two-phase sampling can be described as follows:

[Step 1] From the finite population, select a first-phase sample A_1 of size n_1 and observe $\mathbf{x}$.

[Step 2] Treat the first-phase sample A_1 as the population and select a second-phase sample A_2 of size n_2. In this case, the selection probability for the second-phase sample is often determined by the value of $\mathbf{x}$ obtained from the first-phase sample.

Since the second-phase sample selection probability depends on the observed value of the first-phase sample, the sample inclusion probability for the second-phase sample is a random variable in the sense that its value is changed as the first-phase sample changes. In this case, the theory for HT estimation does not

directly applicable. To discuss it further, note that the overall first-order inclusion probability is

$$\pi_i = Pr(i \in A_2) = \sum_{A_2;\ i \in A_2} P(A_2)$$

where

$$P(A_2) = \sum_{A_1; A_2 \subset A_1} P_2(A_2 \mid A_1) P_1(A_1),$$

$P_1(\cdot)$ is the sample selection probability for the first-phase sample and $P_2(\cdot \mid A_1)$ is the sample selection probability for the second-phase sample which is conditional on the first-phase sample. Thus,

$$\begin{aligned}
\pi_i &= \sum_{\{A_2;\ i \in A_2\}} \sum_{\{A_1; A_2 \subset A_1\}} P_2(A_2 \mid A_1) P_1(A_1) \\
&= \sum_{\{A_1;\ i \in A_1\}} \sum_{\{A_2; A_2 \subset A_1 \& i \in A_2\}} P_2(A_2 \mid A_1) P_1(A_1) \\
&= \sum_{\{A_1;\ i \in A_1\}} \sum_{\{A_2;\ i \in A_2\}} P_2(A_2 \mid A_1) P_1(A_1).
\end{aligned}$$

If we define the conditional first-order inclusion probability for the second-phase sampling as

$$\pi^{(2)}_{i|A_1} = Pr(i \in A_2 \mid A_1)$$

then the first order inclusion probability is

$$\begin{aligned}
\pi_i &= \sum_{A_1;\ i \in A_1} \pi^{(2)}_{i|A_1} P_1(A_1) \qquad (11.1) \\
&= E_1\left(\pi^{(2)}_{i|A_1}\right),
\end{aligned}$$

where $E_1(\cdot)$ is the expectation with respect to the first-phase sampling. Here, the conditional first-order inclusion probability $\pi^{(2)}_{i|A_1}$ is a random variable in the sense that it is a function of $\mathbf{x}$ in the first phase sample A_1. The conditional expectation (11.1) cannot be computed because we have only one realization of A_1.

If the sampling design satisfies the invariance condition as in two-stage sampling, then $\pi^{(2)}_{i|A_1} = \pi^{(2)}_i$ and, by (11.1), we have

$$\begin{aligned}
\pi_i &= \sum_{A_1;\ i \in A_1} P_1(A_1) \pi^{(2)}_i \\
&= \pi^{(1)}_i \pi^{(2)}_i.
\end{aligned}$$

In this case, HT estimator can be implemented.

To discuss unbiased estimation for two-phase sampling, first consider the following quantity

$$\hat{Y}_1 = \sum_{i \in A_1} \frac{y_i}{\pi_i^{(1)}}$$

which is unbiased for the population total of y. Thus, we have only to construct an unbiased estimator of $\hat{Y}_1$ from the two-phase sample. Using the HT estimation idea conditionally, we obtain

$$\hat{Y}^* = \sum_{i \in A_2} \frac{y_i}{\pi_i^{(1)} \pi_{i|A_1}^{(2)}}. \tag{11.2}$$

If we define $\pi_i^* = \pi_i^{(1)} \pi_{i|A_1}^{(2)}$ then (11.2) can be written as $\hat{Y}^* = \sum_{i \in A_2} y_i / \pi_i^*$. Thus, its conditionally unbiased estimator (11.2) is sometimes called π^*-estimator.

The π^*-estimator is conditional unbiased to $\hat{Y}_1$, which is unbiased to Y. Thus, it is unbiased unconditionally. Also, the variance is

$$V\left(\hat{Y}^*\right) \quad = \quad V\left\{E\left(\hat{Y}^* \mid A_1\right)\right\} + E\left\{V\left(\hat{Y}^* \mid A_1\right)\right\}$$

which leads to

$$V\left(\hat{Y}^*\right) = V\left\{\sum_{i \in A_1} \frac{y_i}{\pi_i^{(1)}}\right\} + E\left\{\sum_{i \in A_1} \sum_{j \in A_1} \left(\pi_{ij|A_1}^{(2)} - \pi_{i|A_1}^{(2)} \pi_{j|A_1}^{(2)}\right) \frac{y_i}{\pi_i^*} \frac{y_j}{\pi_j^*}\right\}. \tag{11.3}$$

Here, $\pi_{ij|A_1}^{(2)} = Pr\left(i \in A_2, j \in A_2 \mid A_1\right)$ is the conditional joint inclusion probability. The variance decomposition in (11.3) has two parts where the first part is the variance due to first-phase sampling and the second part is the variance due to second-phase sampling.

11.2 Two-phase sampling for stratification

Stratified sampling is one of the most popular method of sampling design that improve the efficiency of the point estimator. To apply stratified sampling, stratification variable should be available in the finite population. If it is not the case, one may use the two-phase sampling approach. Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{iH})$ be the vector of

stratification variables where x_{ih} takes the value one if unit i belongs to stratum h and takes the value zero otherwise. The auxiliary variable $\mathbf{x}$ is not available in the finite population.

The two-phase sampling for stratification can be described as follows:

1. Perform a SRS of size n from the finite population and obtain $\sum_{i\in A_1}\mathbf{x}_i = (n_1, n_2, \cdots, n_H)$ where $n = \sum_{h=1}^{H} n_h$.

2. Among the n_h elements, select r_h elements by SRS independently across the strata, where r_h is determined after when the first-phase sample is selected.

In this two-phase sampling, the realized sample size n_h in stratum h is a random variable and follows the multinomial distribution approximately as

$$(n_1, n_2, \cdots, n_H) \sim MN(n; W_1, W_2, \cdots, W_H)$$

where $MN(n;\mathbf{p})$ denotes the multinomial distribution with parameter $\mathbf{p}$ and $W_h = N_h/N$ is the population proportion of stratum h.

In this two-phase sampling, the π^*-estimator of the population mean of y is

$$\hat{\bar{Y}}_{tp} = \sum_{h=1}^{H} w_h \bar{y}_{h2} \tag{11.4}$$

where $w_h = n_h/n$ and $\bar{y}_{h2} = r_h^{-1}\sum_{i\in A_2} x_{ih} y_i$. Since the expectation of $w_h = n_h/n$ is $W_h = N_h/N$, the π^*-estimator in (11.4) can be viewed as the stratified sample estimator when the stratum size W_h is unknown. The total variance is, by (11.3), obtained as

$$V\left(\hat{\bar{Y}}_{tp}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 + E\left\{\sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2 \left(\frac{1}{r_h} - \frac{1}{n_h}\right) s_{h1}^2\right\} \tag{11.5}$$

where

$$s_{h1}^2 = \frac{1}{n_h - 1}\sum_{i\in A_1} x_{ih}\left(y_i - \bar{y}_{h1}\right)^2$$

and $\bar{y}_{h1} = n_h^{-1}\sum_{i\in A_1} x_{ih} y_i$.

Also, if we define $\bar{y}_1 = n^{-1}\sum_{h=1}^{H} n_h \bar{y}_{h1}$ and

$$s_1^2 = \frac{1}{n-1}\sum_{i\in A_1}(y_i - \bar{y}_1)^2$$

then we have $E\left(s_1^2\right) = S^2$ and

$$s_1^2 = \sum_{h=1}^{H}\left\{\frac{n_h}{n-1}(\bar{y}_{h1} - \bar{y}_1)^2 + \frac{n_h - 1}{n-1}s_{h1}^2\right\}.$$

Thus, (11.5) is approximately equal to

$$V\left(\hat{\bar{Y}}_{tp}\right) = E\left\{n^{-1}\sum_{h=1}^{H} w_h(\bar{y}_{h1} - \bar{y}_1)^2 + \sum_{h=1}^{H} r_h^{-1} w_h^2 s_{h1}^2\right\}. \tag{11.6}$$

Here, the finite population correction term is ignored. The variance formula in (11.6) is expressed as a sum of the two terms. One is a function of between-stratum variance and the other is a function of within-stratum variance. In computing the between-stratum variance, we used n observations, while, in computing the within-stratum variance, we used r_h observations in each stratum. Thus, if the between-stratum variance is larger than the within-stratum variance, the efficiency of the two-phase estimator is increased.

Also, the variance formula in (11.6) gives an idea for variance estimation. Since (11.6) is expressed as an expectation of the quantities that can be computed from the first-phase sample, we can use $\bar{y}_{h2}$ and s_{h2}^2 instead of $\bar{y}_{h1}$ and s_{h1}^2, respectively, in (11.6). Thus,

$$\hat{V}\left(\hat{\bar{Y}}_{tp}\right) = E\left\{n^{-1}\sum_{h=1}^{H} w_h\left(\bar{y}_{h2} - \hat{\bar{Y}}_{tp}\right)^2 + \sum_{h=1}^{H} r_h^{-1} w_h^2 s_{h2}^2\right\} \tag{11.7}$$

is an approximately unbiased estimator of the variance in (11.6). Instead of (11.7), we can also use jackknife method to estimate the variance of the two-phase sampling estimator. See Kim, Navarro, and Fuller (2006) for details.

To compare the variance (11.6) of the two-phase sampling estimator with that of the simple random sampling estimator of equal size $r = \sum_{h=1}^{H} r_h$, note that

$$V\left(\hat{\bar{Y}}_{SRS}\right) = E\left\{r^{-1}\sum_{h=1}^{H} w_h(\bar{y}_{h1} - \bar{y}_1)^2 + r^{-1}\sum_{h=1}^{H} w_h s_{h1}^2\right\}$$

and so

$$V\left(\hat{\bar{Y}}_{SRS}\right) - V\left(\hat{\bar{Y}}_{tp}\right) = E\left\{\left(\frac{1}{r}-\frac{1}{n}\right)\sum_{h=1}^{H} w_h\left(\bar{y}_{h1}-\bar{y}_1\right)^2 + \sum_{h=1}^{H}\left(\frac{1}{r}-\frac{w_h}{r_h}\right)w_h s_{h1}^2\right\}.$$

The first term is always positive as $r < n$ and the second term is zero under proportional allocation ($r_h = rw_h$) but it can be made positive for optimal allocation.

To discuss optimal allocation, first consider the cost function. The total cost can be expressed as

$$C = c_1 n + \sum_{h=1}^{H} c_{2h} r_h$$

and, writing $\nu_h = r_h/n_h$, the optimal allocation can be determined by minimizing

$$V = \frac{1}{n}\left\{\left(S^2 - \sum_{h=1}^{H} W_h S_h^2\right) + \sum_{h=1}^{H} W_h S_h^2 \frac{1}{\nu_h}\right\}$$

subject to

$$C = n\left(c_1 + \sum_{h=1}^{H} c_{2h} W_h \nu_h\right).$$

To find the optimal allocation, we have only to find the set of ν_h's that minimizes $V \times C$. The optimal solution is

$$\nu_h^* = \left(\frac{c_1}{c_{2h}} \times \frac{S_h^2}{S^2 - \sum_{h=1}^{H} W_h S_h^2}\right)^{1/2} \tag{11.8}$$

which lead to

$$\frac{r_h^*}{n^*} = W_h \nu_h^* = W_h \left(\frac{c_1}{c_{2h}}\right)^{1/2}\left(\frac{S_h^2}{S^2 - \sum_{h=1}^{H} W_h S_h^2}\right)^{1/2}. \tag{11.9}$$

If $c_{2h} = c_2$ and $S_h^2 = S_w^2$ for all h, then

$$\frac{r^*}{n^*} = \left(\frac{c_1}{c_2}\right)^{1/2}\left(\frac{1}{\phi - 1}\right)^{1/2} \tag{11.10}$$

where

$$\phi = \frac{S^2}{S_w^2}$$

denotes the relative efficiency due to stratification (under proportional allocation). If $c_2 = 10c_1$ and $\phi = 2$ then $r/n = \sqrt{0.1} = 0.32$.

11.3 Regression estimator for two-phase sampling

In the previous section, the auxiliary variable **x** obtained from the first phase sample is used to design the sampling mechanism for the second-phase sampling. In this section, we consider the case when the auxiliary variable is used at the estimation stage.

To illustrate the idea, assume that the first-phase sample is a simple random sample of size n and the second-phase sample is also a simple random sample of size r from the first-phase sample. Also, since we observe $\mathbf{x}_i$ in the first-phase sample, we compute

$$\bar{\mathbf{x}}_1 = \frac{1}{n}\sum_{i\in A_1}\mathbf{x}_i$$

from the first-phase sample and compute

$$(\bar{\mathbf{x}}_2, \bar{y}_2) = \frac{1}{r}\sum_{i\in A_2}(\mathbf{x}_i, y_i)$$

from the second-phase sample. The two-phase regression estimator is now computed as

$$\bar{y}_{reg,tp} = \bar{y}_2 + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\hat{B} \tag{11.11}$$

where

$$\hat{B} = \left(\sum_{i\in A_2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2}\mathbf{x}_i y_i.$$

Since we can show that $\hat{B} - B = O_p(r^{-1/2})$, we have

$$\bar{y}_{reg,tp} = \bar{y}_2 + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'B + O_p\left(r^{-1}\right)$$

and we can obtain

$$V\left(\bar{y}_{reg,tp}\right) \doteq B'V\left(\bar{\mathbf{x}}_1\right)B + V\left(\bar{e}_2\right) \tag{11.12}$$

where

$$\bar{e}_2 = \frac{1}{r}\sum_{i\in A_2}\left(y_i - \mathbf{x}_i'B\right).$$

Therefore, under the SRS under both phases,

$$V\left(\bar{y}_{reg,tp}\right) \doteq \left(\frac{1}{n} - \frac{1}{N}\right)B'S_{xx}B + \left(\frac{1}{r} - \frac{1}{N}\right)S_{ee} \tag{11.13}$$

where

$$\begin{aligned} S_{xx} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}_N\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}_N\right)' \\ S_{ee} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(e_i-\bar{e}_N\right)^2. \end{aligned}$$

The first term in (11.14) is the variance term explained by $\bar{\mathbf{x}}_1$ and the second term in (11.14) is the variance of $\bar{y}_{reg,tp}$ that is not explained by x. Note that the efficiency gain due to two-phase sampling can be computed as

$$V\left(\bar{y}_2\right)-V\left(\bar{y}_{reg,tp}\right) = \left(\frac{1}{r}-\frac{1}{n}\right)B'S_{xx}B.$$

Thus, when the linear regression relationship between y and $\mathbf{x}$ is strong and r/n is small, the gain of efficiency is higher.

For variance estimation, we can use (11.14) and estimate each component from the sample. That is, we may use

$$\hat{V}\left(\bar{y}_{reg,tp}\right) \doteq \left(\frac{1}{n}-\frac{1}{N}\right)\hat{B}'\hat{S}_{xx,1}B+\left(\frac{1}{r}-\frac{1}{N}\right)\hat{S}_{ee,2} \tag{11.14}$$

where

$$\begin{aligned} \hat{S}_{xx,1} &= \frac{1}{n-1}\sum_{i\in A_1}\left(\mathbf{x}_i-\bar{\mathbf{x}}_1\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}_1\right)' \\ \hat{S}_{ee,2} &= \frac{1}{r-1}\sum_{i\in A_2}\left(y_i-\mathbf{x}_i'\hat{B}\right)^2. \end{aligned}$$

If jackknife is used, one should take into account of the sampling variability of $\bar{\mathbf{x}}_1$ in the regression estimator. See Sitter (1997) and Fuller (1998) for details.

11.4 Repeated surveys

Repeated surveys means that the survey measurement are taken for the same population at different times. For example, in the US Current Population Survey, the employment rates are announced in the monthly basis. In this case, the sample selection can be repeated over different times. In the repeated surveys, suppose that there are two different years, there are three different parameters of interest.

1. $\bar{Y}_1 - \bar{Y}_2$: the difference of the population mean over two different years.

2. $(\bar{Y}_1 + \bar{Y}_2)/2$: overall mean over the two different years.

3. $\bar{Y}_2$: the population mean at the most recent year.

The optimal sampling design for $\theta_1 = \bar{Y}_1 - \bar{Y}_2$ can be quite different from the optimal sampling design for $\bar{Y} = (\bar{Y}_1 + \bar{Y}_2)/2$. Let $\bar{y}_1$ and $\bar{y}_2$ be an unbiased estimator of $\bar{Y}_1$ and $\bar{Y}_2$, respectively. If we use $\hat{\theta} = \bar{y}_1 - \bar{y}_2$ to estimate $\bar{Y}_1 - \bar{Y}_2$, the variance of $\hat{\theta}$ is minimized when $Corr(\bar{y}_1, \bar{y}_2) = 1$. The correlation is increase when the sample for $t = 1$ is the same as the sample for $t = 2$. That is, it is the case when the same sample is used to obtain the measurement for $t = 1$ and for $t = 2$. Such sample is often called panel sample. On the other hand, if the parameter of interest is $\bar{Y} = (\bar{Y}_1 + \bar{Y}_2)/2$, the panel sample design increases the variance. The independent sample selection, leading to $Corr(\bar{y}_1, \bar{y}_2) = 0$, is more efficient than the panel sample design if we are interested in estimating $\bar{Y} = (\bar{Y}_1 + \bar{Y}_2)/2$.

Now, if we are interested in estimating $\bar{Y}_2$, the following partial overlap sampling design is more efficient than the previous two sampling designs.

1. At $t = 1$, using a SRS of size n to obtain A_1.

2. At $t = 2$, first stratify the finite population into two strata. One is A_1 and the other is $U - A_1$. From the first stratum A_1, select a SRS of size n_m to get A_{2m}. From the second stratum $U - A_1$, select a SRS of size $n_u = n - n_m$, independently from the first stratum, to get A_{2u}. The final sample is $A_2 = A_{2m} \cup A_{2u}$. The first stratum is called "matched" stratum and the second stratum is called "unmatched" stratum.

In this case, the final sample in the matched stratum can be viewed as a two-phase sample where the first phase sample is A_1 and the second phase sample is A_{2m}. Also, the final sample in the unmatched sample is also a two-phase stratified sample, where the first phase sample is $U - A_1$ and the second-phase sample is A_{2u}. The following table presents a summary of the two estimators in each two-phase sample.

Stratum	Population Size	Sample Size	Estimator of $\bar{Y}$
Matched	n	n_m	$\hat{\bar{Y}}_m$
Unmatched	$N-n$	n_u	$\hat{\bar{Y}}_u$
	N	n	$\alpha\hat{\bar{Y}}_u+(1-\alpha)\hat{\bar{Y}}_m$

Now, consider the following class of estimators indexed by a constant α:

$$\hat{\bar{Y}}_\alpha = \alpha\hat{\bar{Y}}_u + (1-\alpha)\hat{\bar{Y}}_m. \tag{11.15}$$

Such estimator is a weighted average of two unbiased estimators and often called composite estimator. The composite estimator is (approximately) unbiased if the two components, $\hat{\bar{Y}}_u$ and $\hat{\bar{Y}}_m$, are (approximately) unbiased. The optimal value of α that minimizes the variance of the composite estimator is

$$\alpha^* = \frac{V\left(\hat{\bar{Y}}_m\right) - Cov\left(\hat{\bar{Y}}_u, \hat{\bar{Y}}_m\right)}{V\left(\hat{\bar{Y}}_u\right) + V\left(\hat{\bar{Y}}_m\right) - 2Cov\left(\hat{\bar{Y}}_u, \hat{\bar{Y}}_m\right)}. \tag{11.16}$$

In this case, the optimal composite estimator is

$$\hat{\bar{Y}}^*_\alpha = \alpha^*\hat{\bar{Y}}_u + (1-\alpha^*)\hat{\bar{Y}}_m$$

and its variance is

$$V\left(\hat{\bar{Y}}^*_\alpha\right) = \frac{V\left(\hat{\bar{Y}}_m\right)V\left(\hat{\bar{Y}}_u\right) - \left\{Cov\left(\hat{\bar{Y}}_u, \hat{\bar{Y}}_m\right)\right\}^2}{V\left(\hat{\bar{Y}}_u\right) + V\left(\hat{\bar{Y}}_m\right) - 2Cov\left(\hat{\bar{Y}}_u, \hat{\bar{Y}}_m\right)}. \tag{11.17}$$

To discuss on the choice of unbiased estimators, we first note that the measurement at $t=1$ can be treated as the auxiliary variable x and the measurement at $t=2$ can be treated as the study variable y. In the unmatched stratum, there is no auxiliary information and so we use

$$\hat{\bar{Y}}_u = \frac{1}{n_u}\sum_{i\in A_{2u}} y_i \equiv \bar{y}_u.$$

On the other hand, in the matched stratum, we can use auxiliary information to get

$$\hat{\bar{Y}}_m = \bar{y}_m + (\bar{x}_1 - \bar{x}_m)\, b$$

where

$$\begin{aligned}(\bar{x}_m, \bar{y}_m) &= n_m^{-1} \sum_{i \in A_{2m}} (x_i, y_i) \\ b &= \left\{ \sum_{i \in A_{2m}} (x_i - \bar{x}_m)^2 \right\}^{-1} \sum_{i \in A_{2m}} (x_i - \bar{x}_m) y_i.\end{aligned}$$

Thus, the following summary can be made to the two estimators.

Stratum	Estimator of $\bar{Y}$	Expectation	Variance
Matched	$\hat{\bar{Y}}_m = \bar{y}_m + (\bar{x}_1 - \bar{x}_m) b$	$\bar{Y} + O\left(n^{-1}\right)$	$n_m^{-1}\left(1-\rho^2\right) S^2 + n^{-1}\rho^2 S^2$
Unmatched	$\hat{\bar{Y}}_u = \bar{y}_u$	$\bar{Y}$	$n_m^{-1} S^2$

Also, we can show that

$$Cov\left(\hat{\bar{Y}}_m, \hat{\bar{Y}}_u\right) = 0. \tag{11.18}$$

Thus, the optimal solution (11.16) is

$$\alpha^* = \frac{n n_u - n_u^2 \rho^2}{n^2 - n_u^2 \rho^2}$$

which is equal to $\alpha^* = n_u/n$ for $\rho = 0$ and equal to $\alpha^* = n_u/(n + n_u)$ for $\rho = 1$.

The variance of the optimal composite estimator is then, by (11.17),

$$V\left(\hat{\bar{Y}}_\alpha^*\right) = \frac{n - n_u \rho^2}{n^2 - n_u^2 \rho^2} S^2 \geq \frac{1}{n} S^2 \tag{11.19}$$

with the equality holds when $n_m = n$ or $n_m = 0$ for $\rho \neq 0$.

The optimal allocation minimizing the variance (11.19) is

$$\frac{n_u}{n} = \frac{1}{1 + \sqrt{1 - \rho^2}}, \qquad \frac{n_m}{n} = \frac{\sqrt{1 - \rho^2}}{1 + \sqrt{1 - \rho^2}}.$$

If ρ is large then more sample is selected for the matched stratum. Under this optimal allocation, the variance (11.19) reduces to

$$V\left(\hat{\bar{Y}}_\alpha^*\right) = \frac{S^2}{2n}\left(1 + \sqrt{1 - \rho^2}\right) \tag{11.20}$$

which takes the value between $S^2/(2n)$ and S^2/n. More discuss on this type of repeated surveys can be found in Fuller (1990).

Reference

Fuller, W. A. (1990). Analysis of repeated surveys, *Survey Methodology* **16**, 167-180.

Fuller, W. A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica* **8**, 1153-1164.

Kim, J.K., Navarro, A., and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling, *Journal of the American Statistical Association* **101**, 312-320.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling, *Journal of the American Statistical Association* **92**, 780-787.