Estimation of Variance Components for Multi-Phase Sampling

Jae Kwang Kim
Westat
1650 Research Blvd
Rockville, MD U.S.A. 20850
KimJ@westat.com

Alfredo Navarro
Bureau of Census
Washington DC U.S.A. 20233
alfredo.navarro@census.gov

Wayne A. Fuller
Iowa State University, Department of Statistics
Ames, IA U.S.A. 50011
waf@iastate.edu

We consider estimation and variance estimation for two-phase samples. Let the finite population be of size *N*, indexed from 1 to *N*, and let the finite population be partitioned into *H* strata. Let the parameter of interest be the population total $Y = \sum_{i=1}^{N} y_i$, where y_i is the study variable 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 $\hat{Y}_1 = \sum_{i \in A_1} w_i y_i$, where $w_i = [\Pr(i \in A_1)]^{-1}$, is an unbiased estimator of *Y*. Instead of directly observing y_i for $i \in A_1$, we observe $\mathbf{x}_i = (x_{i1}, ..., x_{iH})$ for all $i \in A_1$ where x_{ih} takes the value one if unit *i* belongs to the *h*-th stratum and is zero otherwise. Assume that $\sum_{h=1}^{H} x_{ih} = 1$. Let $n_h = \sum_{i \in A_1} x_{ih}$ be the number of first phase sample elements in stratum *h* and $r_h = \sum_{i \in A_2} x_{ih}$ be the number of second phase sample elements in stratum *h*, where the second phase sample is selected by stratified random sampling.

Kim et. al. (2000) write the two-phase stratified estimator as

$$\hat{Y}_2 = \sum_{h=1}^{H} \sum_{i \in A_2} x_{ih} w_i \left(\frac{\sum_{s \in A_1} x_{sh} w_s q_s}{\sum_{s \in A_2} x_{sh} w_s q_s} \right) y_i , \tag{1}$$

where $q_i = w_i^{-1}$ for the direct expansion estimator (DEE) and $q_i = 1$ for the reweighted expansion estimator (REE). See Kott and Stukel (1997) for definitions of DEE and REE. For DEE or REE, the variance of the two-phase estimator can be decomposed as

$$Var(\hat{Y}_2) = Var(\hat{Y}_1) + E\left\{ Var(\hat{Y}_2) \mid A_1) \right\} . \tag{2}$$

Kim et. al. (2000) show that the replication variance estimator

$$\hat{V}(\hat{Y}_2) = \sum_{k=1}^{L} c_k (\hat{Y}_2^{(k)} - \hat{Y}_2)^2, \tag{3}$$

where $\hat{Y}_2^{(k)}$ is the k-th version of $\hat{Y}_2$ based on the observations included in the k-th replicate, L is the number of replications, c_k is a factor associated with replicate k determined by the replication method, and

$$\hat{Y}_2^{(k)} = \sum_{h=1}^{H} \sum_{i \in A_2} x_{ih} w_i^{(k)} \left(\frac{\sum_{s \in A_1} x_{sh} w_s^{(k)} q_s}{\sum_{s \in A_2} x_{sh} w_s^{(k)} q_s} \right) y_i ,$$

is a consistent variance estimator for the total variance of $\hat{Y}_2$. Also See Rao and Shao (1992).

To estimate the components in (2), note that the second term, the conditional variance term, can be estimated because the conditional distribution is the distribution generated by stratified random sampling. Thus, a standard replication method treating the second phase stratified sample as an original stratified sample can be used to estimate the conditional variance. The difference between the overall variance estimator of (3) and the conditional variance estimator is a consistent variance estimator for $Var(\hat{Y}_1)$. The estimator uses only elements of the second phase sample.

The estimation method is being applied at the U.S. Census Bureau to data from the Accuracy and Coverage Evaluation conducted in conjunction with the U.S. 2000 Census of Population.

REFERENCES

Kim, J.K., Navarro, A., and Fuller, W.A. (2000). Variance estimation for 2000 census coverage estimates. *American Statistical Association Proc. of the Section on Survey Research Methods.* To appear.

Kott, P.S. and Stukel, D.M. (1997). Can the Jackknife be used with a two-phase sample? *Survey Methodology*, **23**, 81–89.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79**, 811-822.

RESUME

Nous considerons l'estimation de la variance pour l'échantillonnage en deux phases, avec stratification de la deuxieme phase. Nous proposons un estimateur consistant pour la portion de la variance causée par la deuxieme phase d'échantillonnage.