

## MSE Estimation of Homogeneous Linear Estimator in Stratified Multi-stage Sampling

Kyuseong Kim, The University of Seoul

Department of Computer Science and Statistics, Seoul, 130-743, Republic of Korea

**Key Words:** Non-negative definite; Unbiased estimation.

### 1. Introduction

This article is concerned with mean square error (MSE) estimation for the population total estimator under stratified multi-stage sampling. A version of Homogeneous linear estimator (HLE) will be considered as a general total estimator under the design, since a class of HLE includes well known Horvitz-Thompson estimator, ratio and regression estimator as well as Murthy estimator as special cases. In order to make inferences such as confidence interval about HLE, suitable MSE estimator should be found under the sampling design.

Two criteria commonly used in MSE estimation of HLE are unbiasedness and non-negativity. Under a fixed size design in single-stage sampling, the variance estimator proposed Yates and Grundy (1953) of Horvitz-Thomson estimator (HTE), which is a special case of HLE, has a symmetric form and is unbiased, but not always non-negative. Its non-negativity is satisfied only when some conditions hold. Vijayan (1975) find some conditions for non-negative unbiased variance estimation of HTE. By Rao and Vijayan (1977), this method is extended to ratio estimator, which is not HTE but HLE, under sampling with probability proportional to aggregate size and non-negativity conditions are found. More general results are obtained by Vijayan, et. al. (1995) in case of quadratic forms in finite population sampling. They found a necessary form of non-negative unbiased estimator of non-negative definite quadratic form.

In two-stage sampling or more general multi-stage sampling, a method for unbiased variance estimation of HTE is first proposed by Durbin (1953) and extended to HLE by Raj (1966) and Rao (1975). This method is to estimate the first stage variance component as well as the second stage variance component separately and combine two estimators to be unbiased. These methods are confined to unbiased variance estimation, so does not include MSE estimation as well as non-negative estimation. Furthermore, this method could be available only when second stage variance component is estimable.

A different method, called sub-sampling method,

is proposed by Srinath and Hidioglou (1980), where HTE is used for population total estimation. In this method, they make a sub-selection from the original sample in the second stage and construct an estimator by only first stage variance component estimator. The size of sub-selection is determined for the proposed estimator to be unbiased for the total variance. Extending sub-sampling method to HLE, Arnab (1988) proposed an unbiased variance estimator of HLE. Sub-sampling method can be usefully employed when the second variance component may not be estimated. Also this method can be extended to more general sampling procedure not only unbiased HLE but also biased HLE.

In this article, we will present non-negative unbiased MSE estimation methods of homogeneous linear estimator under stratified multi-stage sampling. One is derived in case that second stage variance component is estimable and the other is found when those are not estimable.

### 2. MSE of Homogeneous Linear Estimator

A population is stratified with  $L$  strata with  $N_h$  clusters in the  $h$ th stratum. In the  $h$ th stratum a sample of clusters,  $s_h^*$ , with size  $n_h (\geq 2)$  is selected by a fixed size design  $p_h^*(\cdot)$ , which is selected independently across the strata. In the second stage, some units are drawn from the  $(hi)$ th cluster independently across the clusters by a suitable design. Subsequent sampling can be continued from the selected units independently and we let  $p_{hi}(\cdot)$  denote resulting sampling design in cluster  $(hi)$ .

Let  $Y_{hi}$  denote  $(hi)$ th cluster total and we assume unbiased estimator  $t_{hi}$  of  $Y_{hi}$  is available with finite variance  $V_{h2}(t_{hi}|s_h^*) = \sigma_{hi}^2$  from the selected  $(hi)$ th cluster. Then a version of HLE in stratified multi-stage sampling can be represented as

$$t = \sum_{h=1}^L t_h = \sum_{h=1}^L \sum_{i \in s_h^*} w_{hi}(s_h^*) t_{hi} \quad (1)$$

where  $w_{hi}(s_h^*)$  is a known constant which may depends on either sampling unit  $(hi)$  or sample  $s_h^*$ .

Lots of estimators known to survey statistician are included in the class of HLE. The first example is unit-based HLE such that the coefficient depends only on the unit, i.e.,  $w_{hi}(s_h^*) = w_{hi}$ . If the

unit based HLE is unbiased, it becomes well-known Horvitz-Thompson estimator,

$$t_A = \sum_{h=1}^L \sum_{i \in s_h^*} \frac{t_{hi}}{\pi_{hi}}$$

where  $\pi_{hi}$  is inclusion probability of unit  $(hi)$ . Sample-based HLE is next example, where  $w_{hi}(s_h^*) = w(s_h^*)$ . This includes ratio estimator,

$$t_B = \sum_{h=1}^L \frac{\bar{y}_h(s_h^*)}{\bar{x}_h(s_h^*)} X_h$$

where  $\bar{y}_h(s_h^*)$ ,  $\bar{x}_h(s_h^*)$  are stratum sample means and  $X_h$  the  $h$ th stratum total of variate  $x$ . Third example is unit and sample based HLE, which includes Murthy estimator of the form

$$t_C = \sum_{h=1}^L \sum_{i \in s_h^*} \frac{p(s_h^*|i)t_{hi}}{p(s_h^*)}$$

where  $p(s_h^*|i)$  is selection probability of  $s_h^*$  given unit  $i$ . This estimator depends on both unit  $(hi)$  and sample  $s_h^*$ .

To find MSE of HLE, we may use the following formula,

$$MSE(t) = \sum_{h=1}^L \{M_{h1}[E_{h2}(t|s_h^*)] + E_{h1}[V_{h2}(t|s_h^*)]\} \quad (2)$$

where  $M_{h1}$  and  $E_{h1}$  are calculated over  $p_h^*(\cdot)$  and  $V_{h2}$  and  $E_{h2}$  over  $p_{hi}(\cdot)$ . From the following calculation

$$\begin{aligned} (i) \quad M_{h1}[E_{h2}(t|s_h^*)] &= \sum_{i=1}^{N_h} (b_{hii} - 2b_{hi} + 1)Y_{hi}^2 \\ &+ \sum_{i \neq j}^{N_h} (b_{hij} - b_{hi} - b_{hj} + 1)Y_{hi}Y_{hj} \end{aligned}$$

and

$$\begin{aligned} (ii) \quad E_{h1}[V_{h2}(t|s_h^*)] &= E_{h1}\left[\sum_{i \in s_h^*} w_{hi}^2(s_h^*)\sigma_{hi}^2\right] \\ &= \sum_{i=1}^{N_h} b_{hii}\sigma_{hi}^2 \end{aligned}$$

where  $b_{hij} = \sum_{s_h^* \ni i, j} w_{hi}(s_h^*)w_{hj}(s_h^*)p(s_h^*)$  and  $b_{hi} = \sum_{s_h^* \ni i} w_{hi}(s_h^*)p(s_h^*)$ , we get

$$MSE(t) = \sum_{h=1}^L \left\{ \sum_{i, j=1}^{N_h} d_{hij}Y_{hi}Y_{hj} + \sum_{i=1}^{N_h} b_{hii}\sigma_{hi}^2 \right\} \quad (3)$$

where

$$d_{hij} = b_{hij} - b_{hi} - b_{hj} + 1$$

Here  $(s_h^* \ni k)$  means all sample  $s_h^*$  such that it contains unit  $k$ .

For each stratum, the first part of  $MSE(t)$  is non-negative definite quadratic form. Vijayan et.al. (1995) showed that if a non-negative definite quadratic form  $F(\vec{Y}) = \sum_i \sum_j c_{ij}Y_iY_j$  could be zero for  $\vec{Y} = \vec{z} = (z_1, \dots, z_N)'$  then it can be represented as  $F(\vec{Y}) = -2^{-1} \sum_{i=1}^N \sum_{j=1}^N c_{ij}z_i z_j (Y_i/z_i - Y_j/z_j)^2$ . Directly applying their result to (3), we can obtain a symmetric form of MSE of HLE as

$$MSE(t) = \sum_{h=1}^L \{A_h + C_h\} \quad (4)$$

where

$$A_h = -\frac{1}{2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} d_{hij}z_{hi}z_{hj} \left( \frac{Y_{hi}}{z_{hi}} - \frac{Y_{hj}}{z_{hj}} \right)^2$$

and

$$C_h = \sum_{i=1}^{N_h} b_{hii}\sigma_{hi}^2$$

and  $MSE(t) = 0$  for  $\vec{Y}_h = \vec{z}_h = (z_{h1}, z_{h2}, \dots, z_{hN_h})'$ ,  $h = 1, \dots, L$ . In addition, HLE  $t$  is unbiased for the population total when  $b_{hi} = 1$  for all  $(hi)$ , so its variance can be reduced to

$$Var(t) = \sum_{h=1}^L \{B_h + C_h\}$$

where

$$B_h = -\frac{1}{2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \tilde{d}_{hij}z_{hi}z_{hj} \left( \frac{Y_{hi}}{z_{hi}} - \frac{Y_{hj}}{z_{hj}} \right)^2$$

where  $\tilde{d}_{hij} = b_{hij} - 1$  for all  $(hi)$ .

### 3. MSE Estimation of Homogeneous Linear Estimator

In this section, we investigate non-negative unbiased estimation method for  $MSE(t)$  given in (4). Sampling design  $p_h^*(\cdot)$  in the first stage is already assumed as fixed size design  $p_{hi}(\cdot)$  and subsequent sampling design is assumed to be available as long as it gives unbiased estimator of cluster total. Here, we think two cases either the second stage variance component  $\sigma_{hi}^2$  can be unbiasedly estimated or not.

First, we assume that it is possible to find an unbiased variance estimator of  $t_{hi}$  over  $p_{hi}(\cdot)$ , which

is denoted by  $\widehat{\sigma}_{hi}^2$ . Now we consider a non-negative unbiased MSE estimator of HLE given in (1) such as

$$m_1(t) = \sum_{h=1}^L \left\{ \sum_{i,j \in s_h^*} c_{ij}(s_h^*) t_{hi} t_{hj} + \sum_{i \in s_h^*} e_i(s_h^*) \widehat{\sigma}_{hi}^2 \right\} \quad (5)$$

where  $c_{ij}(s_h^*)$  and  $e_i(s_h^*)$  are known constants depending on either sampling units  $(hi)$  and  $(hj)$  or sample  $s_h^*$  and to be determined later in order to satisfy non-negativity and unbiasedness. Since  $f_h(\vec{Y}_h) = \sum_{i \in s_h^*} \sum_{j \in s_h^*} c_{ij}(s_h^*) t_{hi} t_{hj}$  is non-negative and unbiased for  $F_h(\vec{Y}_h) = \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} d_{hij} Y_{hi} Y_{hj}$  for which  $F_h(\vec{z}_h) = \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} d_{hij} z_{hi} z_{hj} = 0$ , so  $f_h(\vec{z}_h) = 0$ . By applying Vijayan et. al.'s result again, we obtain a symmetric form of  $m_1(t)$  as

$$m_1(t) = \sum_{h=1}^L \{D_h + F_h\} \quad (6)$$

where

$$D_h = -\frac{1}{2} \sum_{i \in s_h^*} \sum_{j \in s_h^*} c_{ij}(s_h^*) z_{hi} z_{hj} \left( \frac{t_{hi}}{z_{hi}} - \frac{t_{hj}}{z_{hj}} \right)^2$$

and

$$F_h = \sum_{i \in s_h^*} e_i(s_h^*) \widehat{\sigma}_{hi}^2$$

Now we find the properties of  $c_{ij}(s_h^*)$  and  $e_i(s_h^*)$ . From (i)  $E_{h2}\{\widehat{\sigma}_{hi}^2\} = \sigma_{hi}^2$ , (ii)  $E_{h2}\{t_{hi}^2\} = \sigma_{hi}^2 + Y_{hi}^2$  and (iii)  $d_{hii} z_{hii} = -\sum_{j(\neq i)} d_{hij} z_{hi} z_{hj}$ , the expectation of  $m_1(t)$  becomes

$$E\{m_1(t)\} = \sum_{h=1}^L \{E\{D_h\} + E\{F_h\}\}$$

where

$$E\{D_h\} = -\frac{1}{2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \left\{ \sum_{s_h^* \ni (i \neq j)} c_{ij}(s_h^*) p(s_h^*) \right\} \phi_{hij}$$

$$\phi_{hij} = z_{hi} z_{hj} \left( \frac{Y_{hi}}{z_{hi}} - \frac{Y_{hj}}{z_{hj}} \right)^2$$

and

$$E\{F_h\} = \sum_{i=1}^{N_h} \left\{ \sum_{s_h^* \ni i} [c_{ii}(s_h^*) + e_i(s_h^*)] p(s_h^*) \right\} \sigma_{hi}^2$$

From equating  $E\{m_1(t)\}$  to  $MSE(t)$ , we can obtain necessary conditions of non-negative unbiased MSE estimator of the form in (6) as

$$(i) E_{h1}\{c_{ij}(s_h^*)\} = \sum_{s_h^* \ni (i \neq j)} c_{ij}(s_h^*) p(s_h^*) = d_{hij} \text{ for all } (hij)$$

and

$$(ii) E_{h1}\{e_{hi}(s_h^*)\} = \sum_{s_h^* \ni i} e_i(s_h^*) p(s_h^*) = 2b_{hi} - 1 \text{ for all } (hi)$$

In addition, if a non-negative unbiased variance estimator has the form in (6), then has the same form as  $m_1(t)$  with different conditions on  $c_{ij}(s_h^*)$  and  $e_i(s_h^*)$  when  $b_{hi} = 1$ .

As candidates for  $c_{ij}(s_h^*)$  and  $e_i(s_h^*)$  we consider some natural choices. For unit-based HLE,  $c_{ij}^{(1)} = d_{hij}/\pi_{hij}$  and  $e_i^{(1)}(s_h^*) = (2b_{hi} - 1)/\pi_{hi}$  can be considered.  $c_{ij}^{(2)}(s_h^*) = d_{hij}/(M_{h2} p_h^*(s_h^*))$  and  $e_i^{(2)}(s_h^*) = (2b_{hi} - 1)/(M_{h1} p_h^*(s_h^*))$  for sample based HLE and  $c_{ij}^{(3)}(s_h^*) = d_{hij} p_h^*(s_h^* | i, j) / p_h^*(s_h^*)$  and  $e_i^{(3)}(s_h^*) = (2b_{hi} - 1) p_h^*(s_h^* | i) / p_h^*(s_h^*)$  for unit and sample based HLE may be candidates. Here  $M_{hi} = \binom{N_h - i}{n_h - 1}$ ,  $i = 1, 2$  and  $n_h$  is sample size of cluster.

When  $\sigma_{hi}^2$  is not estimable, for example  $p_{hi}(\cdot)$  is systematic sampling design, the previous estimator given in (6) is useless. So a different approach for us is needed to estimate MSE of HLE under the situation. Sub-sampling method (Srinath and Hidiroglou, 1980; Arnab, 1988) may well be adapted to the situation. Originally, subsampling method was suggested to find unbiased variance estimator under multi-stage sampling and we will extend the method up to non-negative unbiased MSE estimation under stratified multi-stage sampling.

In the second stage, choose a subsample with less than original sample size by a suitable sampling design and then construct another estimator  $t'_{hi}$  of  $Y_{hi}$  such that  $E_{h2}\{t'_{hi}\}$  and  $V_{h2}\{t'_{hi}\} = \sigma_{hi}^2$ , say. The sub-sample size will be determined later. Now we consider a non-negative unbiased quadratic MSE estimator of the form

$$m_2(t) = \sum_{h=1}^L \left\{ \sum_{i \in s_h^*} \sum_{j \in s_h^*} c_{ij}(s_h^*) t'_{hi} t'_{hj} \right\} \quad (7)$$

As like in (6), we get  $\sum_{i \in s_h^*} \sum_{j \in s_h^*} c_{ij}(s_h^*) z_{hi} z_{hj} = 0$ . So  $m_2(t)$  can be expressed as

$$m_2(t) = \sum_{h=1}^L \left\{ -\frac{1}{2} \sum_{i \in s_h^*} \sum_{j \in s_h^*} c_{ij}(s_h^*) \phi'_{hij} \right\} \quad (8)$$

where

$$\phi'_{hij} = z_{hi}z_{hj} \left( \frac{t'_{hi}}{z_{hi}} - \frac{t'_{hj}}{z_{hj}} \right)^2$$

From the fact that (i)  $E_{h2}\{t'_{hi}\} = Y_{hi}$ , (ii)  $E_{h2}\{t'^2_{hi}\} = \sigma'^2_{hi} + Y^2_{hi}$  and (iii)  $d_{hii}z^2_{hii} = -\sum_{j(\neq i)} d_{hij}z_{hi}z_{hj}$ , we have

$$E\{m_2(t)\} = \sum_{h=1}^L \{G_h + H_h\} \quad (9)$$

where

$$G_h = -\frac{1}{2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \left\{ \sum_{s^*_h \ni (i \neq j)} c_{ij}(s^*_h)p(s^*_h) \right\} \phi_{hij},$$

and

$$H_h = \sum_{i=1}^{N_h} \left\{ \sum_{s^*_h \ni i} c_{ii}(s^*_h)p(s^*_h) \right\} \sigma'^2_{hi}$$

Therefore, equating  $E\{m_2(t)\}$  to  $MSE(t)$  gives the following necessary conditions

$$\begin{aligned} (i) E_{h1}\{c_{ij}(s^*_h)\} &= \sum_{s^*_h \ni (i \neq j)} c_{ij}(s^*_h)p(s^*_h) \\ &= b_{hij} - b_{hi} - b_{hj} + 1 \end{aligned}$$

and

$$(ii) \sigma'^2_{hi} = \frac{b_{hii}}{b_{hij} - b_{hi} - b_{hj} + 1} \sigma^2_{hi} \quad (10)$$

Any design satisfying the equation in (3.9) is available as a sub-sampling design.

Furthermore, a non-negative unbiased variance estimator is of the same form as  $m_2(t)$  with different conditions such that  $b_{hi} = 1$  in (10), which is the same as that of Arnab(1988).

#### 4. CONCLUDING REMARKS

In this paper, we proposed two kinds of MSE estimators of HLE under stratified multi-stage sampling, which are necessary forms of non-negative unbiased MSE estimators. In large scale sample designs where number of strata is large and number of clusters in each stratum is relatively small, the proposed MSE estimators can be effectively used since unequal probability and without replacement sampling design could be employed.

So far, we have thought of the problem about univariate survey variable. Now turning to multivariate survey variables, especially a function of population totals, let  $\theta = \theta(\vec{Y})$  be a function of population totals where  $\vec{Y} = (Y_1, \dots, Y_p)'$ . In bivariate case, includes correlation coefficient, ratio and regression

coefficients and so on. As an usual estimator is plug-in estimator may be considered as  $\hat{\theta} = \theta(t)$  where  $t$  is homogeneous linear estimator. Primary interest here lies in MSE estimation of  $\hat{\theta}$  under stratified multi-stage sampling. The study of the problem is left for further research.

#### References

- Arnab, R (1988). Variance estimation in multi-stage sampling. *The Australian Journal of Stat.*, **30**, 107-110.
- Raj, D. (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*. **61**, 391-396.
- Rao, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhya*, **C37**, 133-139.
- Rao, J.N.K. (1988). Variance estimation in sample surveys. *Handbook of Statistics*, **Vol. 6**, 427-447.
- Rao, J.N.K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Journal of the American Statistical Association*, **72**, 579-584.
- Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231-241.
- Srinath, K.P. and Hidirolou, M.A. (1980). Estimation of variance in multi-stage sampling. *Metrika*, **27**, 121-125.
- Vijayan, K. (1975). On estimating the variance in unequal probability sampling. *Journal of the American Statistical Association*, **70**, 713-716.
- Vijayan, K., Mukhopadhyay, P. and Bhattachayya, S. (1995). On non-negative unbiased estimation of quadratic forms in finite population sampling. *The Australian Journal of Statistics*, **37**, 169-178.
- Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society*, **B15**, 253-261.