Proceedings of the 2013 International Conference on Information, Operations Management and Statistics (ICIOMS2013), Kuala Lumpur, Malaysia, September 1-3, 2013

Estimation of Variance of Horvitz-Thompson Estimator in the presence of Measurement error: The linear Model Approach

Pulakesh Maiti
Indian Statistical Institute, Kolkata .

Abstract

The importance of identifying and finding ways and means to control the non-sampling errors has long been recognized [Mahalanobis (1946); Hansen etal (1951), (1961), Bailer and Dalenius (1970), Dalenious (1974)]. But not much work compared to controlling sampling errors has progressed so far. Attempts have been made here for estimating the two components of the variance, namely sampling variance and measurement variance under the linear model approach. Primary need to estimate measurement variance is to obtain repeat measurements and to follow the principle of randomization i.e., an appropriate survey design needs to be developed. In this paper, a survey design based on Symmetric Balanced Incomplete Block Design (SBIBD) has been developed and deployed in estimating non-sampling variance component of the Horritz-Thompson H-T estimator. The sample design adopted is that of cluster sampling and the estimator used is the H-T estimator.

KEY WORDS: Non-sampling Variance; Symmetric Balanced Incomplete Block Design (SBIBD); Survey Design; H-T estimator, Linear Model.

1. Introduction:

Considered is the set up of simple i.e., direct response on a quantitative response variable *Y* in the context of a finite labeled population of size *N*. In actual surveys, it so happens that we need investigators and often some supervisors as well. A situation wherein there are possibilities of investigators and/or superviser interventions on the response profile finally received by the data collecting agency is depicted. Of course, these intervention effects may be assumed to be random, having mean zero, non-interactive within and between the two sets of people, designated as data collection people. The problem is to estimate the variance of H-T estimator of finite population total of the response variable *Y*, by in corporating a fixed size (*n*) sampling design and by administering the sampling design in a situation where in the above the types of random interventions are likely to be present.

1.1 Definitions and Notations:

Denote by *i* a respondent unit in the sample of size *n* and by *S*[*i*], the number of schedule based observations collected on this particular unit. Naturally, S[i] is based on the 'study/survey design' used for this unit in combination with the scheme for involvement of investigators and supervisors. One may write $S[i] = \Sigma\Sigma I[i;(j.k)]$ where I $[i;(j,k)] = 1$ if (j,k) combination of the investigator and supervisor have both worked on a schedule meant for collecting information from ith responding unit. Naturally, for any triplet $[i;(j,k)]$, I$[i;(j,k)] \geq 0$, while $S(i) > 0$ for each responding unit. whenever I $[i:(j,k)] = 1$, Y $[i;(j,k)]$ is the underlying response.

Sampling Design:

To fix ideas, let us take up the following example of a finite population involving *N*=700 respondents, grouped into *M*=70 clusters, each of ten respondents. Next, a random sample of *n*=7 clusters is drawn following the fixed size(*n*) sampling design, say, for example SRSWOR (70, 7) clusters, each of size 10.

Survey/Study Design:

Let there be seven investigators and 2 supervisers engaged in the process of data collection. This design is essentially derived from a symmetric BIBD (7,7,3,3,1) 'developed from the initial set' (1,2,4)

following Boses technique. Since the cluster size are all equal [=10], we will ignore the cluster size effect and treat each one as a composite respondent unit. (CRU).

Table 1.1 Distribution of Interviewer Assignment into Responding Groups:

Responding Groups / Investigators	U_1	U_2	U_3	U_4	U_5	U_6	U_7
1	✔	✔		✔			
2		✔	✔		✔		
3			✔	✔		✔	
4				✔	✔		✔
5	✔				✔	✔	
6		✔				✔	✔
7	✔		✔				✔

Let there be two supervisers S_1 and S_2 ; The investigators work are assigned randomly in to two supervisers according to the following design.

Table 1.2 Distribution of Investigator's work into two Supervisers:

Investigators / Supervisers	1	2	3	4	5	6	7
S_1	✔	✔	✔	✔			
S_2				✔	✔	✔	✔

Thus, we have choices of $[i:(j,k)]$ for which $I[i:(j,k)]=1$

(I): $(j=1;\ k=1);\ (j=5;\ k=2);(j=7;\ k=2);$

(II): $(j=1;\ k=1);\ (j=2;\ k=1);(j=6;\ k=2);$

(III): $(j=2;\ k=1);\ (j=3;\ k=1);(j=7;\ k=2);$

(IV): $(j=1;\ k=1);\ (j=3;\ k=1);(j=4;\ k=1);\ (j=4;\ k=2);$

(V): $(j=2;\ k=1);\ (j=4;\ k=1);(j=4;\ k=2);\ (j=5;\ k=2);$

(VI): $(j=3;\ k=1);(j=5;\ k=2);\ (j=6;\ k=2);$

(VII): $(j=4;\ k=1);\ (j=4;\ k=2);(j=6;\ k=2);\ (j=7;\ k=2);$

Let us denote by $Y_{[I]}, Y_{[II]}, \ldots, Y_{[VII]}$ the 'data' accrued from the field. It may be noted that each $Y_{[\cdot]}$ is an average of three readings submitted to the agency by the investigator -supervisor combination as dictated by the survey design. It may be noted that each such reading from a CRU is, in effect, the sum of the responses from its constituent 10 responding unit. Without any intervention effect on the part of the investigators/supervisors, these three readings would have been identical for each CRU, as we would have regarded the above data as 'error free' and so usual estimation technique could be routinely used. Naturally this assertion is valid if we further assume that there are no CRU errors.

2. Modeling of the Data.

Our primary objective is to examine the possibility of intervention by one or the other or possibly by both the sources of intervention and thereby provide valid estimate along with standard error. So, we postulate a linear model of the following form, as applied, for example to $Y_{[I;(1,1)]}$

$$Y_{[I;(1,1)]} = TR_{[I]} + IR_1 + S_1 + e_{[II(1,1)]} \qquad (2.1)$$

Where, $TR_{[I]}$ is the true response from CRUI; IR_1 is the intervention effect of the investigator 1 and S_1 is that of the superviser 1; the last term is the so called error term. In case, there are no CRU errors, this term may be dropped. As usual, it is assumed that CRU errors and the intervention effects are all randomly distributed with mean's $0'S$ and variances $\sigma_e^2, \sigma_b^2, \sigma_S^2$ respectively, while all pairwise effects/interventions are uncorrelated.

3. Estimation of the variance of H-T estimator:

We will now discuss essential features of data analysis for unbiased estimation of the finite population total *T*(*Y*) of the study variable and for estimation of the variance of H-T estimator under the above interactive linear model.

In a very general set up, we have a finite labeled population of *N=MK* units, divided into *M* clusters, or CRU'S [as have been termed before] each of size *K*. We take recourse to a fixed size (*n*) sampling design based on the *M* CRUS with positive inclusion probabilities and joint inclusion probabilities of all pairs of CRUS. For example, the well-known sampling design SRSWOR (*M*, *n*) could be utilized.

Because of possible investigator and/or superviser interventions, the response on the study variable Y_i for the CRU labeled *i* may be distorted and we stipulate a model given above. For each CRU *i* in the sample, we simply take the average of the observations underlying it. Under the model assumptions, these serves as an unbiased estimate of the true response TR_i of the i^{th} CRU. i.e., $E_M(Y_{i..}) = TR_i$. Once this ensured, we use the conventional Horvitz-Thompson estimator,

$$T(\hat{TR}) = \sum_{i \in s} Y_{i..} / \pi_i$$

With, $$E(\hat{T}(TR)) = E_S E_M \left(\sum_{i \in s} Y_{i..} / \pi_i \right) = \left(E_S \sum_{i \in s} Y_{i..} / \pi_i \right) = T(TR)$$

An expression for variance of $T(\hat{TR})$ has to be evaluated next.

We use the standard formula,

$$V = V_1 E_2 + E_1 V_2;$$

Here E_2 and V_2 refer to model expectation and model variance. Cleary model expectation results in the true values *TR'S*, And then V_1 refers to computation of the variance of the HTE interms of the *TR's*. Thus for a fixed size (*n*) sampling design,

$$V_1 E_2 = \sum \sum_{i<j} \left[\frac{TR_i}{\pi_i} - \frac{TR_j}{\pi_j} \right]^2 \left(\pi_i \pi_j - \pi_{ij} \right)$$

Next V_2 referes to computation of model variance of the estimator based on the above responses for the CRU'S. The estimator is the HTE for which the model variance involves all individual variances and all pairwise covariances of the averages for the n-sampled CRUS. More explicitly

$$V_M\left(\sum Y_{i..}/\pi_i\right)=\sum_i V_M(Y_{i..})\Big/\pi_i^2+\sum\sum_{i<j} Cov_M\left(Y_{i..},Y_{.j.}\right)\Big/\pi_i\pi_j$$

We now discuss about computation of E_1V_2. It may noted that E_1 refers expectation with respect to the fixed size(n) sampling design. We have assumed $N=MK$ so that all population units are grouped into M cluster or CRUS of size K each. And we also set $m=nk$, so that, in effect, the sampling design corresponds to a fired size (n) sampling design for selection of 'n CRUS each of size k' out of M CRUS in the population, each of size k. Here m represents the total number of ultimate respondents captured in the study. The CRUS are to be regarded as sampling units in our study and the CRU totals are the primary data from each sampled CRU.

Having discussed these 'sampling design issues' we are now in a position to project the concept, underlying E_1.

3.1. Unbiased estimation of variance of estimator:

To find an unbiased estimator of $V_1\ E_2$, a trivial situation would have produced

$$\sum\sum_{i<j}\left[\frac{TR_i}{\pi_i}-\frac{TR_j}{\pi_j}\right]^2\left[\pi_i\pi_j-\pi_{ij}\right]/\pi_{ij},$$

had the TR_i's been known. However in the present situation, TR_i's are unknown and instead we have the unbiased estimates of the TR_i's viz., $Y_{i..}=\hat{T}R_i$. Therefore, one may start with the expression.

$$\sum\sum_{i<j}\left[Y_{i..}/\pi_i-Y_{j..}/\pi_j\right]^2\left[\pi_i\pi_j-\pi_{ij}\right]/\pi_{ij};$$

and work out its expectation i.e., $E_1\,E_2$. It follows that

$$E_2[\ldots]=\sum_{i<j}\left[\frac{TR_i}{\pi_i}-\frac{TR_j}{\pi_j}\right]^2\left[\pi_i\pi_j-\pi_{ij}\right]/\pi_{ij}$$
$$+\sum\sum_{i<j}V_M\left[Y_{i..}/\pi_i-Y_{j..}/\pi_j\right]^2\left[\left(\pi_i\pi_j-\pi_{ij}\right)/\pi_{ij}\right]^2$$

Once more, if we assume the variance components to be known. Then the second term above can be calculated. Hence the first term above can be evaluated by subtraction.

3.2. Illustrative Examples:

Without any loss of generality, we take the sample CRUs to possess the lables (1,2,…,7). We assume that sampling design is SRSWOR (M=70, n= 7). The true population total is $T(TR)$ where each TR is composed of the sum of TR-values of 10 ultimate respondents from within each of the CRUs. We also assume that the reported data correspond to the subtotals based on within CRUs. The true population

subtotals are $TR_1, TR_2, \ldots, TR_{70}$ and our sample of size n=7 provides model based unbiased estimates for $TR_1, TR_2, \ldots, TR_{70}$. Further since we adopt SRSWOR, we assert that

(i) Unbiased estimate of $T(TR)$ *i.e.*, $\hat{T}(TR) = 10 \sum Y_{i..}$

(ii) Unbiased variance estimate is to be computed from

(a) $E_1 V_2$ component: it is just V_2 given by earlier;

(b) $V_2\ E_1$ component: it is the difference between two expressions given by

First expression: $\left[M^2\left(\frac{1}{n}-\frac{1}{M}\right)\right]\left[\sum\sum_{i<j}\left(Y_{i..}-Y_{j..}\right)^2 / \ n(n-1)\right]$;

Second expression: $\left[M^2\left(\frac{1}{n}-\frac{1}{M}\right)\right]\left[(n-1)\sum \sigma_{ii} - \sum\sum_{i<j}\sigma_{ij}\right] \Big/ \ n(n-1)$

It follows, upon simplification that

$$\sum \sigma_{ii} = 25/12\ \sigma_e^2 + 59/24\ \sigma_b^2 + 143/36\ \sigma_S^2;$$

$$\sum\ \sum_{i<j} \sigma_{ij} = 22/9\ \ \sigma_b^2 + 121/36\ \sigma_S^2.$$

By combining the two from (a) and (b) above, we obtain the final expression for unbiased variance estimate as

$$\left[M^2\left(\frac{1}{n}-\frac{1}{M}\right)\right]\left[\sum\sum_{i<j}\left(Y_{i..}-Y_{.j.}\right)^2 / \ n(n-1)\right] \text{ [contribution from data]}$$

PLUS

$$[M/n]\left[\sum_i \sigma_{11}\right] + [M(M-1)/\ n(n-1)]\left[\sum\sum_{i\neq j}\sigma_{ij}\right].$$

the latter expression simplifies to

$$[M/n]\left[25/12\ \sigma_e^2 + 59/24\,\sigma_b^2 + 143/36\ \sigma_S^2\right] +$$
$$[2M(M-1)/n\ (n-1)]\left[22/9\ \sigma_b^2 + 191/36\ \sigma_S^2\right]$$

provided, $\sigma_b^2, \sigma_S^2, \sigma_e^2$ are known.

4. Estimation of $\sigma_b^2, \sigma_S^2, \sigma_e^2$:

Following the methods of data collection and of supervision [Ref. Tables 1.1 and 1.2], there are 3 data points for the first CRU set I as collected by the investigators 1,5,7 for the first CRU. This we do

for all other CRU sets as well. It may be noted that there are altogether 24 data points and the CRU-wise frequency distributions are 3,3,3,4,4,3,4 respectively. We denote by $\underset{\sim}{Y}$ the vector of 24 observations.

4.1. Canonical Reduction of the Data:

Theorem 5.2.1. The $\underset{\sim}{Y}$ can be partitioned as $\underset{\sim}{Y} = \begin{pmatrix} U \\ \sim 7\times 1 \\ V \\ \sim 17\times 1 \end{pmatrix}$ with the dispersion matrix $\sum_{24\times 24}$ which can be partitioned also as $\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$, where, $\sum_{11}$ and $\sum_{22}$ are the variance, covariance matrices of $\underset{\sim}{U}$ and $\underline{V}$.

$$\underset{\sim}{U} = \begin{bmatrix} \frac{1}{\sqrt{3}}(Y_{I11} + Y_{I52} + Y_{I72}) \\ \frac{1}{\sqrt{3}}(Y_{II11} + Y_{II21} + Y_{II62}) \\ \frac{1}{\sqrt{3}}(Y_{III21} + Y_{III31} + Y_{III72}) \\ \frac{1}{\sqrt{4}}(Y_{IV11} + Y_{IV31} + Y_{IV41} + Y_{IV42}) \\ \frac{1}{\sqrt{4}}(Y_{V21} + Y_{V41} + Y_{V42} + Y_{V52}) \\ \frac{1}{\sqrt{3}}(Y_{VI31} + Y_{VI52} + Y_{VI62}) \\ \frac{1}{\sqrt{4}}(Y_{VII41} + Y_{VII42} + Y_{VII62} + Y_{VII72}) \end{bmatrix};$$

$$
\underset{\sim}{V} = \begin{bmatrix}
\frac{1}{\sqrt{2}} y_{I11} - \frac{1}{\sqrt{2}} y_{I52} \\
\frac{1}{\sqrt{6}} y_{I11} + \frac{1}{\sqrt{6}} y_{I52} - \frac{2}{\sqrt{6}} y_{I72} \\
\frac{1}{\sqrt{2}} y_{II11} - \frac{1}{\sqrt{2}} y_{II21} \\
\frac{1}{\sqrt{6}} y_{II11} + \frac{1}{\sqrt{6}} y_{II21} - \frac{2}{\sqrt{6}} y_{II62} \\
\frac{1}{\sqrt{2}} y_{III21} - \frac{1}{\sqrt{2}} y_{III31} \\
\frac{1}{\sqrt{6}} y_{III21} + \frac{1}{\sqrt{6}} y_{III31} - \frac{1}{\sqrt{6}} y_{III72} \\
\frac{1}{\sqrt{2}} y_{IV11} - \frac{1}{\sqrt{2}} y_{IV31} \\
\frac{1}{\sqrt{6}} y_{IV11} + \frac{1}{\sqrt{6}} y_{IV31} - \frac{2}{\sqrt{6}} y_{IV41} \\
\frac{1}{\sqrt{12}} y_{IV11} + \frac{1}{\sqrt{12}} y_{IV31} + \frac{1}{\sqrt{12}} y_{IV41} - \frac{3}{\sqrt{12}} y_{IV22} \\
\frac{1}{\sqrt{2}} y_{V21} - \frac{1}{\sqrt{2}} y_{V41} \\
\frac{1}{\sqrt{6}} y_{V21} + \frac{1}{\sqrt{6}} y_{V41} - \frac{2}{\sqrt{6}} y_{V42} \\
\frac{1}{\sqrt{12}} y_{V21} + \frac{1}{\sqrt{12}} y_{V41} + \frac{1}{\sqrt{12}} y_{V42} - \frac{3}{\sqrt{12}} y_{V52} \\
\frac{1}{\sqrt{2}} y_{VI31} - \frac{1}{\sqrt{2}} y_{VI52} \\
\frac{1}{\sqrt{6}} y_{VI31} + \frac{1}{\sqrt{6}} y_{VI52} - \frac{2}{\sqrt{6}} y_{VI62} \\
\frac{1}{\sqrt{2}} y_{VII41} - \frac{1}{\sqrt{2}} y_{VII42} \\
\frac{1}{\sqrt{6}} y_{VII41} + \frac{1}{\sqrt{6}} y_{VII42} - \frac{2}{\sqrt{6}} y_{VII62} \\
\frac{1}{\sqrt{12}} y_{VII41} + \frac{1}{\sqrt{12}} y_{VII42} + \frac{1}{\sqrt{12}} y_{VII62} - \frac{3}{\sqrt{12}} y_{VII72}
\end{bmatrix}
$$

Proof: The reduction follows immediately by Helmet transformation.

It may be shown that, $\sum_{11}$ Dispersion matrix of $\underset{\sim}{U}$ and $\sum_{22}$ the dispersion matrix of $\underset{\sim}{V}$ can be represented as

$$\sum_{11} = A_1 \sigma_b^2 + A_2 \sigma_s^2 + I \sigma_e^2$$

and (4.1)

$$\sum_{22} = A'_1 \sigma_b^2 + A'_2 \sigma_s^2 + I \sigma_e^2$$

where, A_1, A_2, A'_1, A'_2 are all symmetric matrices of real numbers. After calculation, the matrices A_1, A_2, A'_2, A'_2 in $\sum_{11}$ and $\sum_{22}$ have been found in the tables 4.1, 4.2 and 4.3.

5.4. Estimation of σ_b^2, σ_s^2 and σ_e^2:

5.4.1. Estimation of σ_e^2:

we have, $Y_{Ijk} = TR_I + IR_j + S_k + e_{Ijk}$;

$$(I = 1,2,\ldots,I;\ j = 1,2,\ldots,J \text{ and } k = 1,2,\ldots,k)$$

$$= \overline{TR} + (TR_I - \overline{TR}) + IR_j + S_k + e_{Ijk} \quad (4.2)$$

$$= \overline{TR} + \alpha_i + IR_j + S_k + e_{Ijk}$$

with $\Sigma \alpha_i = 0$.

Let $SSE = \sum_i \sum_j \sum_k \left[y_{Ijk} - y_{I..} - y_{.j.} - y_{..k} + 2y_{...} \right]^2$, then

It can be shown that

$$E(SSE) = \left[IJK - \{(I-1) + (J-1) + (K-1)\} - 1 \right] \sigma_e^2;$$

i.e., $$E\ \{ [SSE/\ [IJK - \{(I-1) + (J-1) + (K-1)\} - 1\ ] \} = \sigma_e^2;$$

Therefore, $$\hat{\sigma}_e^2 = [SSE/\ [IJK - \{I-1) + (J-1) + (K-1)\} - 1] \quad (4.3)$$

5.4.2. Estimation of σ_b^2:

Let ε_1 and λ_1 be the eigen vector corresponding to maximum eigen value λ_1 of the matrix $\underset{\sim}{A_1}$. Now, from (4.1), we have,

$$\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 = \varepsilon_1' A_1 \varepsilon_1 \hat{\sigma}_b^2 + \varepsilon_1' A_2 \varepsilon_1 \hat{\sigma}_b^2 + \hat{\sigma}_e^2 \quad (4.4)$$

After calculaton, all eigen values appeared to be non negative, as they are expected. On computation, maximum eigen value λ_1 becomes 3.5 and $\varepsilon_1' A_2 \varepsilon_1$ becomes – 7.85704 *e* – 005 which is almost 0.

Thus, from (4.1), we have

$$\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 - \hat{\sigma}_e^2 = 3.5\hat{\sigma}_b^2 - .00007857047\hat{\sigma}_s^2$$

Thus, $$\hat{\sigma}_b^2 = \left[\varepsilon_1' \hat{\sum}_{22} \varepsilon_1 - \hat{\sigma}_e^2 \right] \Big/ 3.5 \qquad (4.5)$$

5.4.3. Estimation of σ_s^2:

Method – I:

Simple response variance, i.e., $\left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right)$ can be estimated from the repeat measurements. This can be obtained through our survey design.
Hence,

$$\hat{\sigma}_S^2 = \left(\sigma_b^2 + \sigma_s^2 + \sigma_e^2\right) - \hat{\sigma}_b^2 - \hat{\sigma}_e^2 \qquad (4.6)$$

Method – II:

Let ε_2 and λ_2 be the eigen vector corresponding to maximum eigen value λ_2 of the matrix $\tilde{A}_2$. All the eigen values appeared to be non-negative as they are expected and the maximum eigen value is found to be 1.03332.
Now, from (4.1), we have ,

$$\varepsilon_2' \hat{\sum}_{22} \varepsilon_2 = \varepsilon_2' A_1' \varepsilon_2 \hat{\sigma}_b^2 + \varepsilon_2' A_2' \varepsilon_2 \hat{\sigma}_b^2 + \hat{\sigma}_e^2 \qquad (4.7)$$

$$\varepsilon_2' \hat{\sum}_{22} \varepsilon - 2\hat{\sigma}_e^2 = 1.650562751\hat{\sigma}_b^2 + 1.03332\hat{\sigma}_s^2$$ Now, from (4.3), (4.5) and (4.6), $\hat{\sigma}_S^2$ can be obtained.

References:

1. Bailar, Barbara A. and Tore Dalenius (1970): Estimating the Response Variance Components of the U.S. Bureau of the Census Survey Model, Sankhȳa , 31B, 341 – 360.
2. Dalenius Tore (1974): The Ends and Means of Total Survey Design. Stockholm: The University of Stockholm.
3. Hansen, Morris H. William N. Hurwitz, Etes. Marks, and W. Parker Mauldin (1951): Response Errors in Survey, JASA, 46, 147 – 190.
4. Hansen, Morris H, William N. Hurwitz and Max A. Bershad (1961): Measurement Errors in Survey, JASA, 46, 147 – 190.
5. Mahalanobis, P.C. (1946): Recent Experiments in Statistical Sampling in the Indian Statistical Institute, JRSS, 109, 327 – 328.

Acknowledgement:

The author acknowledges to Professors G.M. Saha and Bikas Sinha, now retired from the Indian Statistical Institute for their valuable suggestions.