TWO-STAGE SAMPLING UNDER MEASUREMENT ERROR

Promod K. Chandhok, Ohio University

It is now recognized that the results obtained from a sample survey are subject to many types of errors (See Dalenius, 1977 a,b,c). In addition to sampling errors, there may be errors of response or measurement, errors due to non-response or non-contact, and errors arising during data analysis. Standard text books in survey sampling, such as Hansen et al. (1953), Cochran (1977), Raj (1968), and Kish (1965) discuss the effect of measurement errors when a simple random sample of units is selected from a population. But modern sample surveys are based on the selection of clusters with unequal probabilities, there being one or more stages of sampling. It is thus important to consider the effect of measurement errors in such situations. The purpose of this paper is to do that. The clusters are selected with unequal probabilities without replacement. In two-stage sampling a simple random sample of subunits is selected from each cluster. The effect of response errors on the bias, variance and the mean square error of the usual estimator is examined.

2. Single Stage Cluster Sampling

Suppose the population comprises N clusters (say, parcels of land) for which auxiliary information $(X_1, X_2, \ldots, X_N)$ on X (say, area) is available from a previous census. A sample of N clusters is selected without replacement with probabilities based on X. Let Π_i be the probability that the cluster U_i is selected in the sample and Π_{ij} the probability that both U_i and U_j are included in the sample. The observation y_{it} made for the character y (say, production of corn) on U_i is subject to measurement error. We shall use the fairly general model

$$y_{it} = Y_i + \mu_i + e_{it} = Y_i' + e_{it} \qquad (1)$$

where Y_i is the true value of Y and μ_i is the bias associated with the cluster U_i. For the errors e_{it} we assume

$$E_2(e_{it}/i) = 0,$$
$$V_2(e_{it}/i) = \sigma^2{}_{ei}, \text{ and}$$
$$\mathrm{Cov}(e_{it}, e_{jt}/i,j) = \rho_o \sigma_{ei}\sigma_{ej}, \; j \neq i$$

Using the usual estimator

$$\hat{Y} = \sum^n \frac{Y_{it}}{\Pi_i} \qquad (2)$$

for estimating $\hat{Y} = \sum^N Y_i$, we have

$$E_2(\hat{Y}) = \sum^n \frac{Y_i'}{\Pi_i} \quad , \; E(\hat{Y}) = \sum^N Y_j' = Y + \mu$$

Thus the bias in $\hat{Y}$ is the sum of the biases associated with the individual clusters in the population. If the measurement errors are not systematic, the bias μ may be small. To find $V(\hat{Y})$, we have (Raj, 1968, p. 52)

$$V_1E_2(\hat{Y}) = \sum^N \frac{1-\Pi i}{\Pi i} Y_i'^2 + \sum_i \sum_{j \neq i} \frac{\Pi_{ij} - \Pi_i\Pi_j}{\Pi_i\Pi_j} Y_i' Y_j' \qquad (3)$$

$$= \sum_i \sum_{j>i} (\Pi_i\Pi_j - \Pi_{ij}) \left(\frac{Y_i'}{\Pi_i} - \frac{Y_j'}{\Pi_j}\right)^2 \qquad (4)$$

$$V_2(\hat{Y}) = \sum^n \frac{\sigma_{ei}{}^2}{\Pi_i^2} + \sum_i^n \sum_{j \neq i} \frac{1}{\Pi_i\Pi_j} \rho_o \sigma_{ei}\sigma_{ej}$$

$$E_1V_2(\hat{Y}) = \sum^N \frac{\sigma_{ei}{}^2}{\Pi_i} + \rho_o \sum_i^N \sum_{j \neq i} \frac{\Pi_{ij}}{\Pi_i\Pi_j} \sigma_{ei}\sigma_{ej} \qquad (5)$$

Hence

$$V(\hat{Y}) = \sum_i \sum_{j>i} (\Pi_i\Pi_j - \Pi ij) \left(\frac{Y_i'}{\Pi_i} - \frac{Y_j'}{\Pi_j}\right) + \sum \frac{\sigma_{ei}{}^2}{\Pi i}$$
$$+ \rho_o \sum_i^N \sum_{j \neq i} \frac{\Pi_{ij}}{\Pi_i\Pi_j} \sigma_{ei}\,\sigma_{ej} \qquad (6)$$

The first term of this expression represents the sampling variance and the second and third terms represent the response variance.

In case $\rho_o \geq 0$ and the individual response biases μ_i are all zero, the variance of $\hat{Y}$ increases when measurement errors are present. If σ_{ei}'s are zero, the response variance vanishes.

The mean square error of $\hat{Y}$, which gives the total error of the estimator, is given by

$$MSE(\hat{Y}) = V(Y) + (\Sigma\mu_i)^2$$

In order to estimate $V(\hat{Y})$ from the sample, we may use (Yates and Grundy, 1953) $\hat{V}(\hat{Y})$ where

$$\hat{V}(\hat{Y}) = \sum_i^n \sum_{j>i} \frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_{ij}} \left(\frac{Y_{it}}{\Pi_i} - \frac{Y_{jt}}{\Pi_j}\right)^2 \qquad (7)$$

Now

$$E_2\hat{V}(\hat{Y}) = \sum_i^n \sum_{j>i} \frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_{ij}} \left(\frac{Y_i'}{\Pi_i} - \frac{Y_j'}{\Pi_j}\right)^2 + \frac{\sigma_{ei}{}^2}{\Pi_i^2} + \frac{\sigma_{ej}{}^2}{\Pi_j^2} - 2\rho_o \frac{\sigma_{ei}\,\sigma_{ej}}{\Pi_i\,\Pi_j}]$$

$$E\hat{V}(\hat{Y}) = E_1 E_2 \hat{V}(\hat{Y}) = \sum_i \sum_{j>i} \frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_{ij}} \left(\frac{Y_i'}{\Pi_i} - \frac{Y_j'}{\Pi_j}\right)^2$$
$$+ \sum \frac{1-\Pi_i}{\Pi_i} \sigma_{ei}{}^2 - \rho_o \sum_i \sum_{j \neq i} \left(\frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_i\Pi_j}\right) \sigma_{ei}\,\sigma_{ej} \qquad (8)$$

$$E\hat{V}(\hat{Y}) = V(\hat{Y}) - \Sigma\sigma_{ei}{}^2 - \rho_o \sum_i \sum_{j \neq i} \sigma_{ei}\,\sigma_{ej}$$

This shows that the customary estimator of the variance understates the true variance of $\hat{Y}$ when $\rho_o \geq 0$. In case $\sigma_{ei}{}^2 = \sigma_{eo}{}^2$, the bias in $\hat{V}(\hat{Y})$ is $-N\sigma_{eo}{}^2[1+(N-1)\rho]$. It may also be noted that the variance estimator $\hat{V}(\hat{Y})$ (Horvitz and Thompson, 1952), where

$$\hat{V}(\hat{Y}) = \sum_i^n \frac{1-\Pi_i}{\Pi_i} Y_{it}{}^2 + \sum_i \sum_{j \neq i} \frac{\Pi_{ij} - \Pi_i\Pi_j}{\Pi_{ij}} \frac{Y_{it}}{\Pi_i} \cdot \frac{Y_{jt}}{\Pi_j} \qquad (9)$$

has the same bias as (7). Furthermore, the bias in both variance estimators is independent of Π_{ij} that determine the probability set-up.

This means that the bias of variance estimation is the same, whether or not sampling is with equal probabilities.

3. Two-Stage Sampling

We shall now consider the situation in which the sample clusters or PSU'S (say,villages) are selected according to the sampling scheme of Sec. 2 and a simple random sample of m_i sub units (say,households) are selected from the M_i sub units listed in the i-th cluster. Denoting the observation made on the kth sub unit belonging to the i-th cluster as y_{ikt}, we assume the model

$$y_{ikt} = Y_{ik} + \mu_{ik} + e_{ikt} = Y'_{ik} + e_{ikt} \qquad (10)$$

where Y_{ik} is the true value of the character Y for the sub unit and μ_{ik} is the bias associated with the sub unit. Furthermore, let $E\,(e_{ikt}/i,k) = 0$, $V(e_{ikt}/i,k) = \sigma_e^2$,

$$\mathrm{Cov}(e_{ikt},e_{jkt}) = 0 \quad i \neq j$$

We shall use the usual estimator (Raj, 1968, p.118)

$$\hat{Y} = \sum_i \frac{M_i}{\Pi_i} \frac{\sum_k y_{ikt}}{m_i} \qquad (11)$$

for estimating the population total Y for y. We have

$$E_3(\hat{Y}) = \sum_i^n \frac{M_i}{\Pi_i} \frac{\sum^n Y'_{ik}}{M_i}, \quad E_2E_3(\hat{Y}) = \sum_i^n \frac{Y'_i}{\Pi_i}$$

$$E(\hat{Y}) = \sum_i^n Y_i = Y + \mu$$

Hence the bias in $\hat{Y}$ for estimating Y is $\mu = \Sigma\mu_i$, the sum of the biases associated with the PSU's in the population. If the measurement errors are not in the same direction, μ may not be high and the bias may be low. In order to calculate the variance of $\hat{Y}$, we use the formula

$$V(\hat{Y}) = E_1E_2V_3(\hat{Y}) + E_1V_2E_3(\hat{Y}) + V_1E_2E_3(\hat{Y})$$

It is easy to see that

$$V_1E_2E_3(\hat{Y}) = \sum_i \sum_{j>i} (\Pi_i\Pi_j - \Pi_{ij})\left(\frac{Y'_i}{\Pi_i} - \frac{Y'_j}{\Pi_j}\right)^2$$

$$V_2E_3(\hat{Y}) = \sum_i^n \frac{M_i^2}{\Pi_i^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{\nu i}^2, \quad S_{\nu i}^2 = \sum_k^{M_i} \frac{(Y'_{ik} - \bar{Y}'_i)^2}{M_i - 1}$$

$$E_1V_2E_3(\hat{Y}) = \sum^N \frac{M_i^2}{\Pi_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{\nu i}^2$$

$$V_3(\hat{Y}) = \sum^n \frac{M_i^2}{\Pi_i^2}\,\frac{\sigma_e^2}{m_i}\,[1+(m_i-1)\rho]$$

$$E_1E_2V_3(\hat{Y}) = \sigma_e^2 \sum_i^N \frac{M_i^2}{\Pi_i}\,\frac{1}{m_i}\,[1+(m_i-1)\rho]$$

Hence

$$V(\hat{Y}) = \sum_i \sum_{j>i} (\Pi_i\Pi_j - \Pi_{ij})\left(\frac{Y'_i}{\Pi_i} - \frac{Y'_j}{\Pi_j}\right)^2 + \Sigma \frac{M_i^2}{\Pi_i} \cdot \left[\frac{1}{m_i} - \frac{1}{M_i}\right]S_{\nu i}^2 + \sigma_e^2 \sum_i^n \frac{M_i^2}{\Pi_i}\,\frac{1}{m'_i}\,[1+(m_i-1)\rho] \qquad (12)$$

The first-term in Eg. (12) represents the between-PSU sampling variance and the second, the within-PSU sampling variance. The third term is the response variance. If there are no biases associated with the sub units and $\rho=0$ (the response deviations are uncorrelated), the additional contribution to the total variance due to measurement errors is $\sigma_e^2 \Sigma M_i^2/(\Pi_i M_i)$. When the sample is self-weighting with k as the overall raising factor, this contribution is $\sigma_e^2 kMo$, where $M_o = \Sigma M_i$ is the total number of sub units in the population. We will now study the status of the variance estimator $\hat{V}(\hat{Y})$ obtained by following the rule developed in Raj (1968, p. 215). This rule gives

$$\hat{V}(\hat{Y}) = \sum_i^n \sum_{j>i} \frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_{ij}}\left(\frac{M_i\bar{y}_i}{\Pi_i} - \frac{M_i\bar{y}_j}{\Pi_j}\right)^2 + \sum_i^n \frac{M_i^2}{\Pi_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{yi}^2 \qquad (13)$$

When there are no measurement errors, $\hat{V}(\hat{Y})$ is known to be unbiased for $\hat{V}(\hat{Y})$ of Eq. (12). Assuming that the measurement errors follow Eq. (10), we have

$$E_3\left(\frac{M_i\bar{y}_i}{\Pi_i} - \frac{M_i\bar{y}_i}{\Pi_j}\right)^2 = \left(\frac{M_i}{\Pi_i}\frac{\sum^n Y'_{ik}}{m_i} - \frac{M_j}{\Pi_j}\frac{\sum^n Y'_{jk}}{m_j}\right)^2 + \sigma_e^2\left[\frac{M_i^2}{\Pi_i^2}\,\frac{1+(m_i-1)\rho}{m_i} + \frac{M_j^2}{\Pi_j^2}\,\frac{1+(m_j-1)\rho}{m_j}\right]$$

$$E_2E_3\left(\frac{M_i\bar{y}_i}{\Pi_i} - \frac{M_j\bar{y}_j}{\Pi_j}\right)^2 = \left(\frac{Y'_i}{\Pi_i} - \frac{Y'_j}{\Pi_j}\right)^2 + \frac{M_i^2}{\Pi_i^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{\nu i}^2 + \frac{M_j^2}{\Pi_j^2}\left(\frac{1}{m_j} - \frac{1}{M_j}\right)S_{\nu j}^2 + \sigma_e^2\left[\frac{M_i^2}{\Pi_i}\,\frac{1+(m_i-1)\rho}{m_i} + \frac{M_j^2}{\Pi_j^2} \cdot \frac{1+(m_j-1)\rho}{m_j}\right]$$

This gives

$$E_1E_2E_3 \sum_i \sum_{j>i} \frac{\Pi_i\Pi_j - \Pi_{ij}}{\Pi_{ij}}\left(\frac{M_i\bar{y}_i}{\Pi_i} - \frac{M_j\bar{y}_j}{\Pi_j}\right)^2$$

$$= V(\hat{Y}) - \Sigma M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{\nu i}^2 - \sigma_e^2\, \Sigma \frac{M_i^2[1+(m_i-1)\rho]}{m_i}$$

Expressing $s_{yi}^2 = \sum_k (y_{ikt} - \bar{y}_i)^2/(m_i-1)$ as the sum of the squares of the differences of pairs y_{ik}, y_{il}, it is easy to find that

$$E \sum^n \frac{M_i^2}{\Pi_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_{yi}^2 = \Sigma M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{\nu i}^2 + \sigma_e^2(1-\rho)\, \Sigma M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)$$

Hence

$$E\hat{V}(\hat{Y}) = V(\hat{Y}) - \sigma_e^2\, \Sigma M_i\,[1+(M_i-1)\rho] \qquad (14)$$

When $\rho > 0$, the variance estimator $V(\hat{Y})$ is negativeTy biased, the amount of bias being independent of Π_{ij}, the probabilities of inclusion of pairs of units in the sample. Whether the PSU's are selected with equal or unequal probabilities, the bias of the variances estimator (13) is the same and it does not decrease when the sample size is increased.

4. Summary

The usual estimators in two-stage sampling are studied when measurement errors are present. An expression for the contribution of measurement error to the total variance has been obtained. The usual estimates of variance that we use in practice are shown to underestimate the true variance. This fact should be taken into account when interpreting the variance estimates.

References

Chandhok , P.K. (1982). A study of the effects of measurement error in survey sampling. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

Cochran, W.G. (1977). Sampling Techniques, 3rd Ed., Wiley, New York.

Dalenius, T. (1977a). Bibliography of Non-Sampling errors in Surveys, I(A-G). International Statistical Review, 45, 71-89.

__________ (1977b). Bibliography of Non-Sampling errors in Surveys, II(H-Q). International Statistical Review, 45, 181-197.

__________ (1977c). Bibliography of Non-Sampling errors in Surveys, III(R-Z). International Statistical Review, 45, 303-317.

Hansen, M.H., et. al. (1953). Sample Survey Methods and Theory, Wiley, New York.

Kish, L. (1965). Survey Sampling, Wiley, New York.

Raj, D. (1968). Sampling Theory, McGraw-Hill, New York.