STRATIFIED SAMPLING UNDER MEASUREMENT ERROR

Promod K. Chandhok, Ohio University
420 Copeland Hall, Athens, OH 45701

KEY WORDS: stratified sampling, measurement error, optimum allocation

1. INTRODUCTION

Measurement errors are present in most actual surveys. An easy way to deal with these is to pretend they do not exist, or if they do, assume their effect is negligible. Another approach is to model error and, utilizing complex sample designs, estimate contributions to variance by measurement error (Fellegi, 1964). But cost and time constraints may not allow us to follow this approach. This paper looks at the effect of measurement error on stratified sampling.

2. STRATIFIED SAMPLING

In this section the notation and model used are described. Then, the effect of measurement error on stratified mean estimator and its variance is ascertained. Further, the optimum allocation is determined and the variance of stratified mean under optimum allocation is obtained. This variance is then compared with the variance when measurement errors are absent. Finally, the bias of the standard estimate of variance is obtained.

2.1 Notation and Model

Let a population of N units be divided into H non-overlapping strata (or subpopulations) of sizes $N_1, N_2, \ldots, N_H$ such that $\sum_{h=1}^{H} N_h = N$. For an estimator

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \hat{\bar{Y}}_h \tag{1}$$

where $\hat{\bar{Y}}_h$ is an estimate of the population mean of the y-values of the units in the h-th stratum and W_h $(h = 1,2,\ldots,H)$ are constants, we have

$$\text{Bias}(\bar{y}_{st}) = \sum_{h=1}^{H} W_h \text{ Bias}(\hat{\bar{Y}}_h) \tag{2}$$

and

$$V(\bar{y}_{st}) = \sum_h W_h^2 V(\hat{\bar{Y}}_h) \tag{3}$$

Equation (3) is obtained by assuming samples are selected independently in each stratum. Consider the model

$$\begin{aligned} y_{hi} &= Y_{hi} + \mu_{hi} + e_{hi} \\ &= Y'_{hi} + e_{hi} \end{aligned} \tag{4}$$

where y_{hi} is the observed value of the (h,i)-th unit (h denotes the stratum and i the unit within the stratum); Y_{hi} the true value of the (h,i)-th unit; and μ_{hi}, e_{hi} the bias and error respectively associated with the (h,i)-th unit. This model will hold when an interviewer enumerates the units of only one stratum and all interviewers assigned to a stratum are similar. For samples of n_h and n_l units from the h-th and l-th strata respectively, we assume

$$\begin{aligned} E(e_{hi}|h,i) &= 0 \\ V(e_{hi}|h,i) &= \sigma_h^2 \\ \text{Cov}(e_{hi},e_{hj}|h,i,j) &= \rho_h\sigma_h^2,\ i \neq j \\ \text{Cov}(e_{hi},e_{\ell j}|h,\ell,i,j) &= 0,\ h \neq \ell \end{aligned} \tag{5}$$

where σ_h^2 is the variation between repeated measurements on any unit in a stratum h and ρ_h, the correlation between measurements on any two units within a stratum. A simple random sample of n_h units is selected without replacement from the h-th stratum. Let

$$\bar{y}_h = n_h^{-1} \sum_i^{n_h} y_{hi}$$

$$\bar{Y}'_h = N_h^{-1} \sum_i^{N_h} Y'_{hi}$$

$$S^2_{Y'h} = (N_h - 1)^{-1} \sum_i^{N_h} (Y'_{hi} - \bar{Y}'_h)^2$$

2.2 Principal Results

For the estimator

$$\bar{y}_{st} = \sum W_h \bar{y}_h$$

we have

$$V(\bar{y}_{st}) = \sum W_h^2\ V(\bar{y}_h)$$

using (3). Under model (4) along with the assumptions (5) we have,

$$V = V(\bar{y}_{st}) = \sum_h W_h^2\{(\frac{1}{n_h} - \frac{1}{N_h})S^2_{Y'h} + \frac{\sigma_h^2}{n_h}[1 + (n_h - 1)\rho_h]\} \tag{6}$$

We now consider the problem of allocation of sample size to strata. The criterion for determining the vector $(n_1,n_2,\ldots,n_H)$ is either to minimize $V(\bar{y}_{st})$ for a fixed cost or to minimize cost for a fixed variance. Let c_h be the cost of collecting information from a unit in stratum h, and c_0 the overhead cost. Then the total cost of the survey is

$$C = c_0 + \sum c_h n_h \tag{7}$$

To determine optimum allocation we shall follow the approach in Raj (1968). The variance of stratified mean is of the form $\Sigma(A_h/n_h)$ where

$$A_h = W_h^2\{S^2_{Y'h} + \sigma_h^2(1 - \rho_h)\}$$

The terms in the variance which are independent of n_h are ignored since they are not pertinent to this problem. Using Cauchy-Schwarz inequality we infer that the product V.C is minimum if and only if

$$n_h \propto W_h\sqrt{\{S^2_{Y'h} + \sigma_h^2(1 - \rho_h)\}} \,/\, \sqrt{c_h}$$

for all h. Hence

$$\frac{n_h}{n} = \frac{W_h\sqrt{\{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)\}/c_h}}{\sum\limits_h^L [W_h\sqrt{\{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)\}/c_h}]} \tag{8}$$

This implies that n_h, the size of sample selected from the h-th stratum, should be larger if, for the h-th stratum:

(1) the Y'-values are more variable, or
(2) cost of sampling is lower, or
(3) size N_h is larger, or
(4) σ^2_h the within-trial variance is large, or
(5) ρ_h the correlation between errors is low.

For the no-measurement error case, i.e. $\sigma_h = 0$ and $\mu_h = 0 \;\forall\; h$, (8) reduces to

$$\frac{n_h}{n} = \frac{W_h S_{Yh}/\sqrt{c_h}}{\sum W_h S_h/\sqrt{c_h}}$$

which is the well known formula for the no-measurement error case.

We now have the allocation of sample size to strata. Suppose the sample is chosen to minimize $V(\bar{y}_{st})$ for specified cost, then on substituting the optimum values of n_h from (8) in the cost function (7), we have

$$n = \frac{(C - c_0)\sum_h W_h\sqrt{\{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)\}/c_h}}{\sum_h W_h\sqrt{c_h\{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)\}}} \tag{9}$$

If V is fixed, then n can be found by substituting the optimum values of n_h in the equation (6).

If $c_h = c$ for $h = 1,2,\ldots,H$, then the cost is

$$C = c_0 + cn$$

and optimum allocation for fixed cost reduces to optimum allocation for fixed n. Then

$$n_h = \frac{nW_h\sqrt{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)}}{\Sigma W_h\sqrt{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)}} \tag{10}$$

This allocation will be called the modified Neyman allocation. Again, when measurement errors are absent, equation (10) reduces to

$$n_h = n\frac{W_h S_{Yh}}{\Sigma W_h S_{Yh}}$$

The minimum value of $V(y_{st})$ when n is fixed is

$$V_{min}(\bar{y}_{st}) = \frac{1}{n}\cdot(\Sigma W_h\sqrt{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)})^2 - \Sigma\frac{W_h^2 S^2_{Y'h}}{N_h} + \Sigma W_h^2\sigma_h^2\rho_h \tag{11}$$

This allocation is optimum when measurement errors are considered. Equation (11), under no-measurement error case, reduces to

$$V_{min}(\text{no-error}) = \frac{1}{n}(\Sigma W_h S_{Yh})^2 - \Sigma\frac{W_h^2 S^2_{Yh}}{N_h} \tag{12}$$

Also, on comparing the measurement error case with the no-measurement error case, we have

$$\begin{aligned}V_{min} - V_{min}(\text{no-error}) = {} & \frac{1}{n}(\Sigma W_h\sqrt{S^2_{Y'h} + \sigma^2_h(1 - \rho_h)})^2 \\ & - \frac{1}{n}(\Sigma W_h S_{Yh})^2 \\ & - \Sigma\frac{W_h^2 S^2_{Yh}}{N_h} + \Sigma\frac{W_h^2 S^2_{Yh}}{N_h} \\ & + \Sigma W_h^2\sigma_h^2\rho_h\end{aligned}$$

Assume $S_{Y'h} = S_{Yh}$, and to simplify expressions let

$S_{mh} = \sqrt{S^2_{Yh} + \sigma^2_h(1 - \rho_h)}$, then

$$\begin{aligned}n\{V_{min} - V_{min}(\text{no-error})\} = {} & (\Sigma W_h S_{mh})^2 \\ & - (\Sigma W_h S_{Yh})^2 + n\Sigma W_h^2\sigma_h^2\rho_h \\ = {} & \Sigma W_h^2[S^2_{Yh} + \sigma^2_h(1 - \rho_h)] \\ & + \sum_{h\neq\ell} W_h W_\ell S_{mh} S_{m\ell} \\ & - \Sigma W_h^2 S^2_{Yh} - \sum_{h\neq\ell} W_h W_\ell S_{Yh} S_{Y\ell} \\ & + n\,\Sigma W_h^2\sigma_h^2\rho_h \\ = {} & \Sigma W_h^2\sigma_h^2\{1 + (n - 1)\rho_h\} \\ & + \sum_{h\neq\ell} W_h W_\ell\{S_{mh}S_{m\ell} - S_{Yh}S_{Y\ell}\}\end{aligned}$$

If $\rho_h \geq -(n - 1)^{-1}$, which is usually the case, then variance under measurement error is greater than variance when measurement error is absent. However, if $\rho_h < -(n - 1)^{-1}$, then variance under measurement error can be smaller than variance when measurement error is not present.

Next, consider the standard estimate of variance of $\bar{y}_{st}$,

$$v(\bar{y}_{st}) = \frac{1}{N^2}\Sigma N_h(N_h - n_h)\frac{s_h^2}{n_h}$$

where

$$s_h^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$$

Using the result for simple random sampling without replacement (Chandhok, 1982)

$$\begin{aligned}E[(1 - f)\{n(n - 1)\}^{-1}\sum_{j=1}^{n}(y_j - \bar{y})^2] &= (1 - f)(nN)^{-1}\Sigma(Y_j - \bar{Y})^2 \\ &\quad - N^{-2}\Sigma\sigma_j^2(1 - \rho) - \rho N^{-2}(\Sigma\sigma_j)^2\end{aligned}$$

We can easily see that

$$E\{v(\bar{y}_{st})\} = V(\bar{y}_{st}) - N^{-2}\Sigma N_h(1 - \rho_h)\sigma_h^2 - N^{-2}\Sigma\rho_h\sigma_h^2$$

If the measurement errors are positively correlated, which is usually the case, then the usual estimator underestimates the variance. Even if measurement errors are uncorrelated, this estimator underestimates the variance. However, if measurement error is negatively correlated, this estimator can overestimate the true variance.

3. REFERENCES

Bailar, B.A., and T. Dalenius. 1969. Estimating Response Variance Components of the U.S. Bureau of the Census Survey Model. Sankhya, Series B, 31:341-360.

Chandhok, P.K. 1982. A Study of the Effects of Measurement Error in Survey Sampling. Unpublished Ph.D. Dissertation, Department of Statistics, Iowa State University, Ames, Iowa.

Chandhok, P.K. 1986. Two Stage Sampling Under Measurement Error. Proc. Survey Res. Methods Sec., Amer. Stat. Assoc., 650-652.

Cochran, W.G. 1977. Sampling Techniques. Wiley, New York.

Fellegi, I.P. 1964. Response Variance and its Estimation. J. Amer. Stat. Assoc. 59:1016-1041.

Hansen, M.H., and B.J. Tepping. 1969. Progress and Problems in Survey Methods and Theory Illustrated by the Work of the United States Bureau of the Census. In N.L. Johnson and H. Smith, Jr. (eds.). New Developments in Survey Sampling. Wiley Interscience, New York.

Hansen, M.H., W.N. Waksberg. 1970. Research on Non-Sampling Errors in Censuses and Surveys. Rev. Int. Stat. Inst. 38:318-32.

Hansen, M.H., W.N. Hurwitz, and M.A. Bershad. 1960. Measurement Errors in Censuses and Surveys. Bull. de Institut. International de Statistique 38(2):359-374.

Pritzker, L., and R. Hanson. 1962. Measurement Errors in the 1960 Census of Population. Proc. Soc. Stat. Section Amer. Stat. Assoc. 80-89.

Raj, D. 1968. Sampling Theory. McGraw Hill, New York.

Sukhatme, P.W., and B.V. Sukhatme. 1970. Sampling Theory of Surveys with Application. Iowa State University Press, Ames, Iowa.