A NOTE ON THE OPTIMUM SAMPLE SIZE WHEN THERE ARE NON-SAMPLING ERRORS

By V. Mukerji

G.I.P.E., Poona

Non-sampling errors can arise on a number of counts, some of them being controllable and some uncontrollable. As examples of controllable non-sampling or measurement errors one could cite those arising out of faulty organization, the lack of care taken in obtaining the observations, the deficiency in the expertise of the personnel, etc. One would expect that such controllable errors could be reduced but at a higher cost per unit of observation. It is difficult, however, to estimate empirically the relationships between cost per unit of observation and the non-sampling errors or some characteristics of their distribution. One could, however, hypothecate some simple plausible relationships and on the basis of these relationships determine the optimum sample size and the cost per unit of observation to see if some broad conclusions regarding the choice of the sample size, etc., which may be operationally useful, can be inferred. With this object in view, very simple and plausible decreasing functional relationships between the means and variances of non-sampling errors on the one hand and cost per unit of observation on the other have been assumed and with the further assumption of a linear cost function, expressions for optimum sample sizes have been obtained for simple random sample designs as an illustration.

2. Let the true variable corresponding to the characteristic or trait under consideration be denoted by y and let the observed variable corresponding to the observation on the characteristic or trait be denoted by y' and let ϵ denote the (additive) non-sampling-error which is the deviation of the observed variable from the true variable.

Then

$$y' = y + \epsilon$$

Let the mean and variance of y be denoted by m and σ_y^2 and those of ϵ by β and σ_ϵ^2. Then the mean and variance of y' are given by

$$E(y') = m + \beta$$

and

$$V(y') = \sigma_y^2 + \sigma_\epsilon^2 + 2\sigma_{y\epsilon}$$
$$= \sigma_y^2 + \sigma_\epsilon^2 \qquad (1)$$

if y and ϵ can be assumed to be uncorrelated as may be true in some cases. β may be called the measurement-bias and ϵ, the response-variance.

It is assumed as indicated earlier that

$$\beta^2 = \frac{k_1}{c_1^{l_1}} \qquad l_1,\ l_2 > 0.$$

and

$$\sigma_\epsilon^2 = \frac{k_2}{c_2^{l_2}}, \qquad (2)$$

where c_1 is the marginal cost per observation, the total cost for n observations being given by

$$C = c_0 + c_1 n \qquad (3)$$

c_0 standing for the fixed costs of the survey. l_1 and l_2 may be called the cost-elasticity coefficients of measurement bias and measurement (response) variance respectively.

Let the population-characteristic that is to be estimated be m, he population mean of y. For a simple random sample the estimate of m is given by

$$\bar{y}' = \frac{\sum_{r=1}^{n} y_r'}{n}, \qquad (4)$$

where y_1', y_2', $\cdots$ y_n' are the n observations in the sample. Here,

$$E(\bar{y}') = m + \beta \neq m, \text{ if } \beta \neq 0.$$

Further

$$V(\bar{y}') = \frac{\sigma_y^2 + \sigma_\epsilon^2}{n} \qquad (5)$$

sampling is done with equal probability with replacement. In all the calculations below, the assumption of sampling with replacement is made. However, the results obtained can be shown to be valid for sampling without replacement also with some minor modifications.

A linear loss-function L defined by

$$L = a_1\beta^2 + a_2\sigma_{\bar{y}'}^2$$

of which the mean square error (M.S.E.) criterion is a particular case with $a_1 = a_2 = 1$, is chosen as the criterion for determining the optimum sample size and the corresponding optimum c_1, the cost per unit of observation. So, (i) if the total cost cannot exceed a fixed quantity C_0, then n and c_1 have to be so chosen as to minimize L subject to $C \leqslant C_0$. (ii) If the loss L cannot exceed a given quantity L_0, then n and c_1 have to be chosen so as to minimize C subject to $L \leqslant L_0$. It can be readily verified that optimum n and c_1 in cases (i) and (ii) are given by

$$\text{(i)} \quad n = \frac{C_0 - c_0}{c_1} \quad \text{with } c_1 \text{ given by}$$

$$a_2\sigma_y^2 c_1^{l_1+l_2+1} + (1 - l_2)\, k_2 a_2 c_1^{l_1+1} - a_1 k_1 l_1 (C_0 - c_0)\, c_1^{l_2} = 0,$$

and

$$\text{(ii)} \quad n = \frac{a_2\left(\sigma_y^2 + \dfrac{k_2}{c_1^{l_2}}\right)}{L_0 - \dfrac{a_1 k_1}{c_1^{l_1}}}$$

with c_1 given by

$$L_0\sigma_y^2 c_1^{l_1+l_2} + L_0 k_2 (1 - l_2)\, c_1^{l_1} - (l_1 + 1)\, a_1 k_1 \sigma_y^2 c_1^{l_1} = a_1 k_1 k_2 (1 - l_2 + l_1).$$

Particular cases of (*a*) $l_2 = 0$ or fixed σ_ϵ^2; (*b*) $l_1 = 0$, $k_1 = 0$ or zero β and (*c*) $a_1 = a_2 = 1$, the M.S.E. criterion may be derived from the above general results. Other criteria such as minimizing C subject to $\beta \leqslant a\sigma_{\bar{y}'}$ for a specified a and $\sigma_{\bar{y}'}^2 \leqslant \sigma_0^2$ or of minimizing $\sigma_{\bar{y}'}^2$ subject to $C \leqslant C_0$ and $\beta \leqslant a\sigma_{\bar{y}'}$ or of minimizing β subject to $C \leqslant C_0$ and $\sigma_{\bar{y}}^2 \leqslant \sigma_0^2$ may also be considered.*

Similar results can be obtained for any sample design for which the variance of the estimator can be expressed as V/n, V being the variance per sample unit.†

---

* Explicit results are obtained in reference[4].

† Results for stratified random sampling with the additional assumptions of that the cost per observation is the same in all strata and β and σ_e^2 are the same in all strata are obtained in reference[4].

Generalization of these results to non-linear cost-functions and other forms of dependence relationships between β and σ_ϵ^2 on the one hand and the cost-function on the other can be considered. It is very difficult if not almost impossible to establish empirically the forms of the latter dependence relationships and hence only simple dependence relationships have been considered here.

3. One broad trend that is discernible in general in results (i) and (ii) is that the larger the σ_y^2 or the more heterogeneous the population (of the true variable y), the larger has n got to be and the smaller has c_1 got to be. Even with all the assumptions made and with the restrictive model this result is valid under certain conditions only. Thus for the case $\beta = 0$ the conclusion is valid only when $l_2 > 1$. When $l_2 \leqslant 1$ maximum efficiency is achieved when n coincides with the population size and $c_1 = 0$ irrespective of heterogeneity or otherwise of the population.

SUMMARY

Non-sampling error has been treated as a random variable whose mean and variance are taken to be decreasing functions of the varying cost per unit of observation. The determination of the optimum sample size and the cost per unit when such non-sampling errors are present has been studied in some situations.

ACKNOWLEDGEMENT

I am very grateful to the referee for suggestions and comments and for drawing my attention to the paper[3] by Hansen, Hurwitz and Bershad.

REFERENCES

1. Sukhatme, P. V. .. *Sampling Theory of Surveys with Applications.*
2. Cochran, W. G. .. *Sampling Techniques.*
3. Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. "Measurement errors in census and surveys," *Bull. of Int. St. Inst.*, 1961, **38** (2), 359–74.
4. Mukerji, V. .. "Size and cost of surveys and observationl errors, —a quantitative model" (unpublished paper, G I.P.E., Poona 4).