Survey Methodology, December 1993
Vol. 19, No. 2, pp. 193–197
Statistics Canada

Sampling from Imperfect Frames with Unknown Amount of Duplication

SHIBDAS BANDYOPADHYAY and A.K. ADHIKARI[1]

ABSTRACT

This study covers such imperfect frames in which no population unit has been excluded from the frame but an unspecified number of population units may have been included in the list an unspecified number of times each with a separate identification. When the availability of auxiliary information on any unit in the imperfect frame is not assumed, it is established that for estimation of a population ratio or a mean, the mean square errors of estimators based on the imperfect frame are less than those based on the perfect frame for simple random sampling when the sampling fractions of perfect and imperfect frames are the same. For estimation of a population total, however, this is not always true. Also, there are situations in which estimators of a ratio, a mean or a total based on smaller sampling fraction from imperfect frame can have smaller mean square error than those based on a larger sampling fraction from the perfect frame.

KEY WORDS: Imperfect frame; Efficiency.

1. INTRODUCTION

A frequent problem that arises while planning surveys is the non-availability of complete frames. The International Statistical Institute recognized the importance of studying the problem of sampling from imperfect frames and arranged discussions by experts on this topic during its 34th Session held in Ottawa, Canada where Hansen *et al.* (1963) and Szameitat and Schaffer (1963) presented invited papers. One may also refer to Singh (1977, 1983). Wright and Tsao (1983) have written a bibliography on frames to bring attention to problems which arise when sampling from imperfect frames.

Recently two separate surveys were undertaken by the Indian Statistical Institute to evaluate the impact of government sponsored programmes for the uplift of economic conditions of fishermen's community in West Bengal, India. In the first survey (1988), the households were selected using the membership registers of the Fishermen's Co-operative Societies (FCS). In the second and more recent survey, the list of beneficiary fishermen of the Fish Farmer's Development Agency (FFDA) was used. It was known that not all FCS members or FFDA beneficiaries would be from different households, but it was not possible to identify the FCS members or the FFDA beneficiaries belonging to the same household without contacting the households. Thus, when FCS membership registers or FFDA beneficiary lists were used for household selection, the frames contained an unknown number of duplication. Since the household information was collected by personal interview, it was possible to identify the duplication in the selected households only. The values of the variables associated with the households in the sample were divided by the respective number of duplications in the frame while retaining the duplicate households in the sample under separate identification.

The set-up of imperfect frames discussed here is a special case of Rao (1968). One of the referees has pointed out that the situation discussed in the paper also occurs at Statistics Canada in certain frames for business surveys.

Imperfect frames to be covered in this study are those in which no population unit has been excluded from the frame but any population unit may have been included in the frame an unspecified number of times with a separate identification each time. It is assumed that it would be possible to ascertain, at the data collection stage, the number of duplicates in the frame for each selected unit. The possibility of selecting two or more duplicates of a population unit in the sample is not excluded. The availability of auxiliary information on the units in the imperfect frame is not assumed and only simple random sampling without replacement (SRSWOR) schemes are discussed.

Since the total number of population units will not be known from the imperfect frames to be covered here, problems of estimation of a mean of a population character and its total are not identical.

Here is the main question discussed in this paper. Which is better: to up-date the imperfect frame and select a sample, or to use the imperfect frame?

In the two surveys on fishermen's households, it was felt that most of the economic variables of interest would be highly related to the number of FCS members/FFDA beneficiaries in a household in the sense that the variability

[1] Shibdas Bandyopadhyay and A.K. Adhikari, Indian Statistical Institute, Calcutta, India 700 035.

of such an economic variable per FCS member/FFDA beneficiary would be less than the variability of the economic variable per household. It was felt that one could effectively use an imperfect frame in such situations.

It will be established that for situations such as above estimators of a ratio, a mean, or a total based on smaller sampling fraction, imperfect frame can have smaller Mean Square Error (MSE) than those based on a larger sampling fraction from the perfect frame.

Even when the variability is not related to the number of duplications as discussed above, it will be established that for estimating a ratio or a mean, using an imperfect frame will be preferable to using a perfect frame, from the MSE point of view, when the sampling fractions of the imperfect and the perfect frames are same.

2. NOTATIONS AND RELATIONS

Consider a finite population consisting of N units $U_1, U_2, \ldots, U_N$. Let $U_1^*, U_2^*, \ldots, U_M^*$ be the units listed in an imperfect frame. For $k = 1, 2, \ldots, r$, let A_k denote the sub-population of the original N units consisting of N_k distinct population units. Each of the units in A_k is listed in the imperfect frame exactly k number of times under separate identifications. Assume that

(a) each U_i belongs to an A_k for some k, (*i.e.*, each U_i is included in the imperfect frame at least once) and

(b) if U_j^* is selected in the sample using the imperfect frame, it will be possible to identify, at the data collection stage, the corresponding U_i and the associated value of k (*i.e.*, the number of duplicates of U_i in the incomplete frame under separate identifications, one of which is the selected unit U_j^*) for which U_i belongs to A_k.

The following relations are valid.

$$N_1 + N_2 + \ldots + N_r = N;$$

$$N_k \geq 0, k = 1, 2, \ldots, r,$$

$$N_1 + 2N_2 + \ldots + rN_r = M,$$

where r, N_1, N_2, ..., N_r, and N are all unknown and only M is known with $M \geq N$; M may be written as, for unknown α,

$$M = N(1 + \alpha), \quad \alpha \geq 0. \tag{2.1}$$

Let X and Y values on the unit U_i be X_i and Y_i respectively, $(i = 1, 2, \ldots, N)$. Since each U_j^*, $(j = 1, 2, \ldots, M)$, can be identified with a U_i for some i, $(i = 1, 2, \ldots, N)$, and since U_i belongs to A_k for some k, $(k = 1, 2, \ldots, r)$, define X, Y and C values for the unit U_j^* as

$$X_j^* = X_i/k, \quad Y_j^* = Y_i/k, \quad C_j^* = 1/k.$$

Because of assumptions (a) and (b), X^*, Y^*, and C^* values are observable for the selected units from the imperfect frame.

The following relations connect the measurements in the imperfect frame to those in the perfect frame.

$$\sum_{j=1}^{M} Y_j^* = M\bar{Y}^* = \sum_{i=1}^{N} Y_i = N\bar{Y};$$

$$\sum_{j=1}^{M} C_j^* = M\bar{C}^* = N;$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)^2 = N\sigma_Y^2 - S(2, Y) + (N\bar{Y})^2(1/N - 1/M),$$

where

$$N\sigma_Z^2 = \sum_{i=1}^{N} (Z_i - \bar{Z})^2$$

and

$$S(a,Z) = \sum_{k=2}^{r} (1 - 1/k)\left\{\sum_{i:U_i \in A_k} Z_i^a\right\}; \tag{2.2}$$

$$\sum_{j=1}^{M} (C_j^* - \bar{C}^*)^2 = N(1 - N/M) - S(0, Y);$$

$$\sum_{j=1}^{M} (Y_j^* - \bar{Y}^*)(C_j^* - \bar{C}^*) = N\bar{Y}(1 - N/M) - S(1, Y).$$

For the unit U_i let

$$D_i = Y_i - \bar{Y}; \; W_i = Y_i - RX_i, \quad \text{where} \quad R = \bar{Y}/\bar{X}. \tag{2.3}$$

Since no auxiliary information on the units is assumed, comparisons will be done on the basis of a SRSWOR sample. Let m be the size of the sample from the imperfect frame and n be the corresponding sample size had the frame been perfect. Define efficiency of a perfect frame compared to the corresponding imperfect frame, for any estimator, as

$$\rho = \frac{\text{MSE based on a sample of size } m \text{ from the imperfect frame}}{\text{MSE based on a sample of size } n \text{ had the frame been perfect}}. \tag{2.4}$$

Also define f as the common sampling fraction when the sampling fractions are same, *i.e.*,

$$n = fN, \quad m = fM = n(1 + \alpha). \tag{2.5}$$

3. RESULTS

Before we proceed to answer the main question raised in Section 1 on the choice of sampling from the perfect frame against sampling from the imperfect frame, we briefly look at the alternatives from cost considerations. If the total cost of up-dating the imperfect frame is expected to be more than the additional cost of data collection from the $(m - n)$ extra units, it is economical to use the imperfect frame with a larger sample size than to update the imperfect frame; this is so when

$$\frac{b_1}{b_0}\left(\frac{m-n}{N}\right) \leq 1, \tag{3.1}$$

where b_1 is the per-unit data collection cost and b_0 is the per-unit up-dating cost. It may be noted that one needs to visit effectively N units to up-date the incomplete frame since the remaining $(M - N)$ units are duplicates and can be identified because of assumption (b). It may also be noted that, even from a SRSWOR sample from the imperfect frame, the extra number of units to be canvassed is at most $(m - n)$ since the sample may contain the same unit under separate identifications. These observations lead to (3.1) for preference of using an imperfect frame.

As has been pointed out in Section 1, the total number of population units N will not be known from the imperfect frame. Thus the problems of estimation of a mean and a total are not identical; the problem of estimation of a mean essentially is the problem of estimation of a ratio, but a total can be estimated directly and unbiasedly, based on a SRSWOR sample of size m from the imperfect frame. It is thus appropriate to estimate a population ratio (similar to domain estimation) with estimation of a mean as a special case, and then to treat estimation of a total separately.

3.1 Estimation of a Ratio

For estimation of a ratio $R = (\bar{Y}/\bar{X})$, the usual ratio estimator is

$$\hat{R} = \bar{y}^*/\bar{x}^*,$$

where the lower case letters represent the corresponding quantities based on a sample, $\bar{y}^*$ is the mean of Y^* values based on a sample of size m from the imperfect frame *etc.* $\bar{y}^*$ and $\bar{x}^*$ are respectively unbiased estimators of $(N\bar{Y}/M)$ and $(N\bar{X}/M)$. Using the delta method the MSE of $\hat{R}$, $E(\hat{R} - R)^2$, is given approximately by

$$\frac{M-m}{m(\bar{X}^*)^2(M-1)M}\sum_{i=1}^{M} W_i^{*2}; \tag{3.2}$$

using the relations of Section 2, (3.2) can be rewritten as

$$\mathrm{MSE}(\hat{R}) = \frac{M(M-m)}{m(N\bar{X})^2(M-1)}\{N\sigma_W^2 - S(2,W)\},$$

where W values are defined in (2.3) and the W^* values correspondingly obtained. It follows from (2.2) that $S(2,W) \geq 0$, and hence from (3.2) one has

$$0 \leq 1 - \frac{S(2,W)}{N\sigma_W^2} \leq 1. \tag{3.3}$$

It now follows from (2.4) that efficiency ρ is

$$\rho = \frac{nM(M-m)(N-1)}{mN(N-n)(M-1)}\left\{1 - \frac{S(2,W)}{N\sigma_W^2}\right\}. \tag{3.4}$$

When sampling fractions are equal, ρ can be written as

$$\rho = \frac{(1+\alpha)(N-1)}{(1+\alpha)(N-1)+\alpha}\left\{1 - \frac{S(2,W)}{N\sigma_W^2}\right\}. \tag{3.5}$$

It, therefore, follows from (3.3) that ρ given by (3.5) satisfies

$$0 \leq \rho \leq 1 \tag{3.6}$$

and thus it is advantageous to use imperfect frame for estimation of a ratio.

It may be noted that $S(2, W)$ is nondecreasing in α and for fixed α, $S(2, W)$ has a larger value when the units with larger W values are replicated in the imperfect frame. Since σ_W^2 is fixed for a given set of N W values, there may be situations in which ρ in (3.4) is less than 1 (as a matter of fact $S(2, W)$ is equal to $N\sigma_w^2$ when W values are all equal and equal to 0) and consequently, there will be situations when sampling from imperfect frame will be preferable even with smaller sampling fraction to sampling from complete frame.

3.2 Estimation of a Mean

As seen in section 3.1, $\bar{y}^*$ is an unbiased estimator of $(N\bar{Y})/M$ where M is known but N is unknown. Thus it is necessary to estimate N to get an estimator for $\bar{Y}$. It may be noted that $\bar{c}^*$ is an unbiased estimator of (N/M), and thus

$$\hat{\bar{Y}} = \bar{y}^*/\bar{c}^*$$

is a natural ratio-type estimator of $\bar{Y}$. On replacing $\bar{x}^*$ in Section 3.1 by $\bar{c}^*$, the MSE of $\hat{\bar{Y}}$ is given by

$$\mathrm{MSE}(\hat{\bar{Y}}) = \frac{M(M-m)}{mN^2(M-1)}\{N\sigma_D^2 - S(2,D)\},$$

where D values are defined in (2.3). Replacing W in Section 3.1 by D we may conclude that (3.6) holds and imperfect frame is better when (2.5) is true.

3.3 Estimation of a Total

To estimate a total, say $N\bar{Y}$, based on a SRSWOR sample of size m from the imperfect frame, the usual estimator is

$$\widehat{(N\bar{Y})} = M\bar{y}^*,$$

which is unbiased for $N\bar{Y}$, with variance

$$\mathrm{MSE}(M\bar{y}^*) = \mathrm{Var}(M\bar{y}^*)$$

$$= \frac{M(M-m)}{m(M-1)}$$

$$\left\{N\sigma_Y^2 - S(2,Y) + (N\bar{Y})^2\left(\frac{1}{N} - \frac{1}{M}\right)\right\}.$$

One may write ρ as

$$\rho = \frac{nM(M-m)(N-1)}{mN(N-n)(M-1)}$$

$$\left\{1 - \frac{S(2,Y) - (N\bar{Y})^2(1/N - 1/M)}{N\sigma_Y^2}\right\}.$$

It is clear from the expression of $\mathrm{Var}(M\bar{y}^*)$ that

$$\left\{S(2,Y) - (N\bar{Y})^2\left(\frac{1}{N} - \frac{1}{M}\right)\right\}\Big/N\sigma_Y^2, \quad (3.7)$$

is less than or equal to unity. However, α and Y values may be so chosen that expression in (3.7) is negative. In such a case, even when (2.5) is true, imperfect frame with larger sampling fraction is inefficient. However, if the scatter of Y^* values are more homogeneous compared to Y values, *i.e.*, if

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \geq \sum_{j=1}^{M}(Y_j^* - \bar{Y}^*)^2, \quad (3.8)$$

then the expression in (3.7) is always nonnegative. Now, one can draw similar conclusions as in Section 3.1, for example, (3.6) is valid when (2.5) is true.

4. AN ILLUSTRATION

As pointed out earlier, in the fishermen's survey, ultimate sampling units of beneficiary-fishermen were selected from the list of beneficiaries available. Being a multidisciplinary survey, many characteristics of the sampling units were observed from each of the sampling unit which either related to the household or to the fishing/fishery enterprise to which the sampling unit belonged. Since only the number of beneficiaries (M) was known and the number of corresponding households/enterprises (N) was not known, it was not possible to see the effect of using the imperfect frame for this survey. However for illustration in this paper, we take the samples drawn from one geographical area (a block within an administrative district in the West Bengal State) as our population and see the effect of resampling from it. In this area, there are 27 beneficiaries (M) and 23 distinct enterprises (N), 19 of the enterprises have single ownership (N_1) and 4 are of joint-ownership type (N_2). Our characteristics of interest are the cost of renovation of water areas (Y) and the acreage of operated water areas (X).

The summary statistics of Y and X are as follows:

$$\sum Y_i = 58{,}815, \quad \sum X_i = 23.36,$$

$$R = \left(\sum Y_i\right)\Big/\left(\sum X_i\right) = 2{,}517.77,$$

$$S(2,Y) = 212{,}201{,}800, \quad S(2,D) = 145{,}101{,}018,$$

$$S(2,W) = 104{,}505{,}327,$$

$$23\sigma_Y^2 = 442{,}702{,}791, \quad 23\sigma_X^2 = 13.6503 \quad \text{and}$$

$$23\sigma_W^2 = 394{,}790{,}716,$$

where W is defined in (2.3).

To find the effect of sampling from the list of 27 beneficiaries we find estimates of

R = Renovation cost per acre of water area,

$\bar{X}$ = Average water area per enterprise in acre and

$N\bar{X}$ = Total acreage of water areas operated by all 23 enterprises.

The table below gives the efficiencies for different choices of m and n.

Efficiency of sampling from perfect frame compared to sampling from imperfect frame (ρ)

Sample sizes		Efficiency for estimators of		
n	m	R	$\bar{X}$	$N\bar{X}$
2	2	0.8695	0.6453	0.9508
4	4	0.8841	0.6561	0.9668
6	6	0.9022	0.6696	0.9866
8	8	0.9225	0.6866	1.0117
8	9	0.7791	0.5781	0.8519
10	10	0.9551	0.7088	1.0444
10	11	0.8172	0.6065	0.8937

It can be seen that in most cases sampling from imperfect frame are more efficient.

ACKNOWLEDGEMENT

Authors wish to thank an Associate Editor and the referees for their valuable suggestions towards improvement of this paper.

REFERENCES

HANSEN, M.H., HURWITZ, W.N., and JABINE, T.N. (1963). The Use of imperfect lists for probability sampling at the U.S. Bureau of Census. *Bulletin of the International Statistical Institute*, 40, 497-517, (with discussions).

INDIAN STATISTICAL INSTITUTE (1988). *A study of Fishermen in West Bengal*: 1985-1986.

RAO, J.N.K. (1968). Some non-response sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

SINGH, R. (1977). A note on the use of incomplete multi-auxiliary information in sample surveys. *Australian Journal of Statistics*, 19, 105-107.

SINGH, R. (1983). On the use of incomplete frames in sample surveys. *Biometrical Journal*, 25, 545-549.

SZAMEITAT, K., and SCHAFFER, K.A. (1963). Imperfect frames in statistics and the consequences for their use in sampling. *Bulletin of the International Statistical Institute*, 40, 517-538, (with discussions).

WRIGHT, T., and TSAO, H.J. (1983). *A frame on frames: An annonated bibliography. Statistical Methods and Improvement of Data Quality*, (Ed. T. Wright). New York: Academic Press, 25-72.