A MULTIPLICITY ESTIMATOR FOR MULTIPLE FRAME SAMPLING

Robert J. Casady and Monroe G. Sirken
National Center for Health Statistics

INTRODUCTION

The counting rule specifies the links between the population elements and the units in the sampling frame at which the units are eligible to be enumerated in a survey [Sirken 1974]. Traditionally, counting rules have been adopted which have the property of uniquely linking population elements to sampling units so that every element is enumerable at one and only one unit. However, the survey design advantages of counting rules which permit the same elements to be enumerated at more than one sampling unit are being increasingly investigated and exploited by network sampling and multiple frame sampling methods.

The theory of multiple frame sampling and the theory of network sampling were developed independently, but they are closely connected, though that connection has not heretofore been investigated. Network sampling may be viewed as a generalization of multiple frame sampling, or alternatively, multiple frame sampling as a special form of stratified network sampling.

From the latter viewpoint, multiple frame sampling assumes a set of frame specific counting rules which permit population elements to be linked to sampling units by different counting rules, but uniquely link every element to one and only one sampling unit by any given counting rule. Stratified network sampling also assumes a set of frame specific counting rules but it does not restrict the number of sampling units linked to a given population element by either the same or different counting rules.

The purpose of this paper is two-fold. First, the definitions and notation which have been utilized in the literature to develop the theory and methods of multiple frame sampling and network sampling are to be consolidated and unified.

Secondly, the multiple frame estimator proposed by Hartley (1962, 1974) is extended to include the situation in which the data for at least one of the sampling frames is collected via a multiplicity counting rule. This generalized Hartley estimator is analytically compared to the stratified sampling multiplicity estimator proposed by Sirken (1972).

A detailed empirical example is presented for a dual frame survey in which it is assumed that the sampling unit for one frame is a household and the sampling unit for the other frame is a telephone. This example is based on data presented by Thornberry and Massey (1978) from the NCHS Household Interview Survey.

It should be noted that all of the work presented in this paper is restricted to the case of two sampling frames so as to avoid unnecessarily complicating the presentation. The extension of the results to the case of more than two frames is straightforward but tedious and unenlightening.

NOTATION AND DEFINITIONS

Let $I = \{I_1, I_2, \ldots, I_N\}$ denote the target population and let $F_1 = \{h_{11}, h_{12}, \ldots, h_{1M_1}\}$ and $F_2 = \{h_{21}, h_{22}, \ldots, h_{2M_2}\}$ denote two sampling frames. The counting rules r_1 and r_2, linking population elements in I to sampling elements in F_1 and F_2 respectively, are represented by the matrices

$[\delta_{ij1}]$ and $[\delta_{ij2}]$ where

$$\delta_{ijk} = \begin{cases} 1 & \text{if } I_j \text{ is linked to } h_{ki}, \quad i=1,2, \\ & \qquad k=1,2 \\ & \qquad j=1,2,\ldots,N \\ 0 & \text{otherwise} \end{cases}$$

The <u>multiplicity</u> of a population element with respect to a particular counting rule is the total number of sampling elements to which it is linked by the rule. Thus, the multiplicity of the population element I_j, with respect to the the counting rule r_k, is given by

$$S_{jk} = \sum_{i=1}^{M_k} \delta_{ijk}, \quad j=1, 2, \ldots, N;\ k=1,2 .$$

The total multiplicity of I_j is given by

$$S_{j\cdot} = S_{j1} + S_{j2}, \quad j=1, 2, \ldots, N .$$

A counting rule is said to be a <u>multiplicity rule</u> if the multiplicity of any population element is greater than one. If no population element has multiplicity greater than one the the rule is said to be a <u>conventional rule</u>.

The <u>coverage set</u> of a <u>counting rule</u> is defined to be the set of population elements which are linked to at least one element in the sampling frame by the counting rule. Thus the coverage set for the counting rule $r_k(k=1,2)$ is given by

$$C_k = \{I_j \mid S_{jk} \geq 1,\ 1 \leq j \leq N\}, \quad k = 1,2 .$$

The <u>total coverage set</u> of the counting rules r_1 and r_2 is $C = C_1 \cup C_2$ and the <u>overlap coverage set</u> is $C_{12} = C_1 \cap C_2$. It will be assumed throughout the remainder of this paper that $C = I$. It should also be noted that the overlap coverage set, C_{12}, is analogous to Hartley's (1974) overlap domain.

In this paper four types of two frame sample survey situations are considered. Specifically, these situations are:

- Type I: $C_{12} = \emptyset$ and both r_1 and r_2 are conventional counting rules,
- Type II: $C_{12} = \emptyset$ and at least one of the counting rules is a multiplicity rule,
- Type III: $C_{12} \neq \emptyset$ and both r_1 and r_2 are conventional counting rules, and
- Type IV: $C_{12} \neq \emptyset$ and at least one of the counting rules is a multiplicity rule.

In the following section it is assumed that probability samples S_1 and S_2 are available from the two frames, however, the two sample survey designs may be entirely different and in fact the two samples may not even be independent.

ESTIMATORS FOR TWO-FRAME SAMPLE SURVEYS

Assume that we want to estimate the aggregate $X = \sum_{j=1}^{N} X_j$ where X_j is the value of the characteristic of interest for the j^{th} element of the target population and let the sample design be specified by the random vectors

$$\underline{a_1} = (a_{11}, a_{12}, \ldots, a_{1M_1})$$

and $$\underline{a_2} = (a_{21}, a_{22}, \ldots, a_{2M_2})$$

where $$a_{ki} = \begin{cases} 1 & \text{if } h_{ki} \varepsilon S_k \\ 0 & \text{otherwise} \end{cases}$$

and $$P_{ki} = P_r \{ a_{ki} = 1 \} > 0$$

$$\text{for } i=1,2,\ldots,M_k,\ k=1,\ 2\ .$$

One of the multiplicity estimators suggested by Sirken (1972) is given by

$$X'_S = \sum_{i=1}^{M_1} (a_{1i}/p_{1i}) \sum_{j=1}^{N} \delta_{ij1} X_j/S_j.$$

$$+ \sum_{i=1}^{M_2} (a_{2i}/p_{2i}) \sum_{j=1}^{N} \delta_{ij2} X_j/S_j.$$

$$= X'_{S1} + X'_{S2} \qquad (1)$$

This estimator is unbiased for X for all four types of two frame surveys.

The multi-frame estimator proposed by Hartley (1964) assumed conventional counting rules and is only appropriate for Type I or Type III sample survey situations. Hartley's estimator can be written as

$$X'_H(\lambda) = \sum_{i=1}^{M_1} (a_{1i}/p_{1i}) \sum_{j\varepsilon B_1-B_2} {}_{ij1} X_j$$

$$+ \sum_{i=1}^{M_2} (a_{2i}/p_{2i}) \sum_{j\varepsilon B_2-B_1} \delta_{ij2} X_j$$

$$+ \lambda \sum_{i=1}^{M_1} (a_{1i}/p_{1i}) \sum_{j\varepsilon B_{12}} \delta_{ij1} X_j$$

$$+ (1-\lambda) \sum_{i=1}^{M_2} (a_{2i}/p_{2i}) \sum_{j\varepsilon B_{12}} \delta_{ij2} X_j$$

$$= X'_{H1} + X'_{H2} + \lambda X'_{H12} + (1-\lambda) X'_{H21} \qquad (2)$$

where B_1, B_2 and B_{12} are index sets defined as

$$B_k = \{j | I_j \varepsilon C_k\}\ k = 1,2$$

and $B_{12} = B_1 \cap B_2$.

Hartley's estimator is unbiased for X in either Type I or Type III sample surveys.

The proposed generalized form of Hartley's estimator to include Type II and Type IV surveys is

$$X'_G(\lambda) = \sum_{i=1}^{M_1} (a_{1i}/p_{1i}) \sum_{j\varepsilon B_1-B_2} \delta_{ij1} X_j/S_{j1}$$

$$+ \sum_{i=1}^{M_2} (a_{2i}/p_{2i}) \sum_{j\varepsilon B_2-B_1} \delta_{ij2} X_j/S_{j2}$$

$$+ \lambda \sum_{i=1}^{M_1} (a_{1i}/p_{1i}) \sum_{j\varepsilon B_{12}} \delta_{ij1} X_j/S_{j1}$$

$$+ (1-\lambda) \sum_{i=1}^{M_2} (a_{2i}/p_{2i}) \sum_{j\varepsilon B_{12}} \delta_{ij2} X_j/S_{j2}$$

$$= X'_{G1} + X'_{G2} + \lambda X'_{G12} + (1-\lambda) X'_{G12}. \qquad (3)$$

It is straightforward to verify that this estimator is unbiased for X for all four types of sample survey situations.

Further it can be verified that the following relationships between the estimators (1), (2) and (3) hold,

Type I: All three estimators are equivalent.

Type II: Sirken's estimator and the generalized Hartley estimator are equivalent.

Type III: The estimators (2) and (3) are equivalent and $X'_S \equiv X'_G(1/2)$. Furthermore the same ancillary information is required for all three estimators.

Type IV: Sirken's estimator and the generalized Hartley estimator are not equivalent and different ancillary information is required for the two estimators. Specifically, for $X'_G(\lambda)$ the multiplicity of each sample element for its frame of enumeration must be known and it must be determined if the element is a member of the overlap coverage set; for X'_S the total multiplicity of each sample element is required.

OPTIMAL ALLOCATION FOR FIXED SURVEY COST

Let W_1 and W_2 be the per unit cost for samples of size m_1 and m_2 from F_1 and F_2 respectively. Assume that the samples are selected independently from the two frames and there exist real numbers α_1, α_2, β_1, β_2, β_{12}, β_{21}, γ_{12} and γ_{21} so that for all values of m_1 and m_2

$$\begin{aligned} \text{Var}\ (X'_{S1}) &= \alpha_1/m_1 \\ \text{Var}\ (X'_{S2}) &= \alpha_2/m_2 \\ \text{Var}\ (X'_{G1}) &= \beta_1/m_1 \\ \text{Var}\ (X'_{G2}) &= \beta_2/m_2 \\ \text{Var}\ (X'_{G12}) &= \beta_{12}/m_1 \\ \text{Var}\ (X'_{G21}) &= \beta_{21}/m_2 \\ \text{Cov}\ (X'_{G1}, X'_{G12}) &= \gamma_{12}/m_1 \\ \text{and}\quad \text{Cov}\ (X'_{G2}, X'_{G21}) &= \gamma_{21}/m_2\ . \end{aligned}$$

Then for a fixed total survey cost W, where $W = W_1 m_1 + W_2 m_2$, the values of m_1 and m_2 that minimize Var (X'_S) are

$$m_1^* = \left(\frac{W}{W_1}\right) \frac{(W_1\alpha_1)^{\frac{1}{2}}}{(W_1\alpha_1)^{\frac{1}{2}} + (W_2\alpha_2)^{\frac{1}{2}}}$$

$$\text{and}\quad m_2^* = \left(\frac{W}{W_2}\right) \frac{(W_2\alpha_2)^{\frac{1}{2}}}{(W_1\alpha_1)^{\frac{1}{2}} + (W_2\alpha_2)^{\frac{1}{2}}}\ .$$

The corresponding minimum variance is

$$\text{Var}_{\min}\ (X'_S) = ((W_1\alpha_1)^{\frac{1}{2}} + (W_2\ \alpha_2)^{\frac{1}{2}})^2/W\ .$$

To minimize Var $(X'_G(\lambda))$ the optimal value of λ, say λ^*, is given implicitly by

$$\frac{W_1^{\frac{1}{2}}\ \phi'_1(\lambda^*)}{\phi_1(\lambda^*)^{\frac{1}{2}}} + \frac{W_2^{\frac{1}{2}}\ \phi'_2(\lambda^*)}{\phi_2(\lambda^*)^{\frac{1}{2}}} = 0$$

where $\phi_1\ (\lambda) = \beta_{12}\ \lambda^2 + 2\ \gamma_{12}\ \lambda + \beta_1$.

$$\phi_2\ (\lambda) = \beta_{21}\ (1-\lambda)^2 + 2\ \gamma_{21}\ (1-\lambda) + \beta_2\ .$$

The optimal values of m_1 and m_2, in terms of λ^*, are given by

$$m_1^* = \left(\frac{W}{W_1}\right)\left(\frac{(W_1\ \phi_1\ (\lambda^*))^{\frac{1}{2}}}{(W_1\ \phi_1\ (\lambda^*))^{\frac{1}{2}} + (W_2\ \phi_2\ (\lambda^*))^{\frac{1}{2}}}\right)$$

$$\text{and}\quad m_2^* = \frac{W}{W_2}\left(\frac{(W_2\ \phi_2\ (\lambda^*))^{\frac{1}{2}}}{(W_1\ \phi_1\ (\lambda^*))^{\frac{1}{2}} + (W_2\ \phi_2\ (\lambda^*))^{\frac{1}{2}}}\right)$$

$$\text{Var}_{\min}\ (X'_G\ \ (\lambda^*)) = ((W_1\ \phi_1\ (\lambda^*))^{\frac{1}{2}} + (W_2\ \phi\ (\lambda^*))^{\frac{1}{2}})^2/W.$$

AN EXAMPLE WITH A HOUSEHOLD FRAME AND A TELEPHONE FRAME

Assume that two frames are available for sampling purposes. The first is a list of household addresses and the second is a list of telephone numbers for telephones located in households. It will be assumed that no more than two telephones are listed for any household. The counting rule for the first frame links persons to the address of the household in which they permanently reside. The counting rule for the second frame links persons to the telephone number of any telephone located in the household in which they permanently reside. It should be noted that the first rule is a conventional counting rule while the second rule is a multiplicity rule since it is assumed that some households have more than one listed telephone. Also, the set of persons linked to a particular household address and the set of persons linked to a listed telephone located at that address correspond.

For the purpose of this example two simplifying assumptions are made:

(1) The coverage set for the household address frame is the entire target population, that is $C_1 = I$.
(2) No household has more than two listed telephone numbers.

Let

M_1 = total number of household addresses in the first frame
M_2 = total number of listed telephones in the second frame
M_3 = total number of households with at least one listed telephone
and M_4 = total number of households with exactly one listed telephone.

Label the household address in the first frame so that

(a) addresses labeled 1 through M_4 correspond to households with only one listed telephone
(b) addresses labeled $M_4 + 1$ through M_3 correspond to households with two listed telephones
(c) addresses labeled $M_3 + 1$ through M_1 correspond to households without any listed telephones.

Then for i=1, 2,..., M_1, $Z_i = \sum_{j=1}^{N} \delta_{ij1}\ X_j$ is the aggregate value of characteristic X for those individuals in the target population who are linked to H_i.

Now suppose we have a SRS, without replacement of m_1 household addresses from the first frame and a SRS, without replacement, of m_2 telephone numbers from the second frame. Letting $\underline{a_1}$ and $\underline{a_2}$ be random vectors such that

$$a_{1i} = \begin{cases} 1 & \text{if the } i^{th} \text{ household is selected} \\ 0 & \text{otherwise} \end{cases}$$

and

$$a_{2i} = \begin{cases} 2 & \text{if two telephone numbers are selected for the } i^{th} \text{ household} \\ 1 & \text{if one telephone number is selected for the } i^{th} \text{ household} \\ 0 & \text{otherwise} \end{cases}$$

The components of the estimators X_S' and $X_G'(\lambda)$ can be written

$$X_{S1}' = (M_1/m_1)\left(1/2 \sum_{i=1}^{M_4} a_{1i} Z_i + 1/3 \sum_{i=M_4+1}^{M_3} a_{1i} Z_i + \sum_{i=M_3+1}^{M_1} a_{1i} Z_i\right)$$

$$X_{S2}' = (M_2/m_2)\left(1/2 \sum_{i=1}^{M_4} a_{2i} Z_i + \sum_{i=M_4+1}^{M} a_{2i} Z_i\right)$$

$$X_{G1}' = (M_1/m_1)\ (\sum_{i=M_3+1}^{M_1} a_{1i} Z_i)$$

$$X_{G2}' \equiv 0$$

$$X_{G12}' = (M_1/m_1) \sum_{i=1}^{M_3} a_{1i} Z_i$$

$$X_{G21}' = (M_2/m_2)\ (\sum_{i=1}^{M_4} a_{2i} Z_i + 1/2 \sum_{i=M_4+1}^{M_3} a_{2i} Z_i)$$

Ignoring finite correction factors, it is straightforward to verify that

$$\mathrm{Var}\ (X_{G1}') = (1/m_1)\left\{M_1(M_1-M_3)\ \sigma^2_{NT} + M_3(M_1-M_3)\ \bar{X}^2_{NT}\right\}$$

$$\mathrm{Var}\ (X_{G12}') = (1/m_1)\left\{M_1\ M_4\ \sigma^2_{1T} + M_1\ (M_3-M_4)\ \sigma^2_{2T} + M_4\ (M_1-M_4)\ \bar{X}^2_{1T} - 2\ M_4(M_3-M_4)\bar{X}_{1T}\ \bar{X}_{2T} + (M_3-M_4)(M_1-M_3+M_4)\bar{X}^2_{21}\right\}$$

$$\mathrm{Var}\ (X_{G21}') = (1/m_2)\ (M_2M_4\ \sigma^2_{1T} + (M_2/2)(M_3-M_4)\ \sigma^2_{2T} + M_4(M_2-M_4)\ \bar{X}^2_{1T} - 2\ M_4(M_3-M_4)\bar{X}_{1T}\ \bar{X}_{2T} + (M_4/2)(M_3-M_4)\bar{X}^2_{2T})$$

$$\mathrm{Cov}(X_{G1}', X_{G12}') = (-1/m_1)(M_1-M_3)(M_4\ \bar{X}_{1T} + (M_3-M_4)\ \bar{X}_{2T})\ \bar{X}_{NT}$$

$$\mathrm{Var}(X_{S1}') = (1/m_1)\left\{(M_1M_4/4)\ \sigma^2_{1T} + (M_1(M_3-M_4)/9\ \sigma^2_{2T} + M_1(M_1-M_3)\ \sigma^2_{NT} + (M_1-M_4)(M_4/4)\ \bar{X}^2_{1T} + ((M_1-M_3+M_4)(M_3-M_4)/9)\ \bar{X}^2_{2T} - (M_4\ (M_3-M_4)/3)\ \bar{X}_{1T}\ \bar{X}_{2T} + M_3(M_1-M_3)\ \bar{X}^2_{NT} - (M_1-M_3)\ M_4\ \bar{X}_{NT}\ \bar{X}_{1T} - 2(M_1-M_3)\{(M_3-M_4)/3\}\bar{X}_{NT}\ \bar{X}_{2T}\right\}$$

$$\mathrm{Var}(X_{S2}') = (1/m_2)\left\{(M_2M_4/4)\ \sigma^2_{1T} + (2M_2\ (M_2-M_4)/9)\sigma^2_{2T} + \{(M_2-M_4)\ M_4/4\}\ \bar{X}^2_{1T} - 2\{M_4(M_3-M_4)/3\}\bar{X}_{1T}\ \bar{X}_{2T} + \{(M_2-M_3-M_4)\}\ \{2(M_3-M_4)\ /9\}\ \bar{X}^2_{2T}\right\}$$

where

$$\bar{X}_{1T} = \sum_{i=1}^{M_4} Z_i/M_4$$

$$\bar{X}_{2T} = \sum_{i=M_4+1}^{M_3} Z_i/(M_3-M_4)$$

$$\bar{X}_{NT} = \sum_{i=M_3+1}^{M_1} Z_i/(M_1-M_3)$$

$$\sigma^2_{1T} = \sum_{i=1}^{M_4}(Z_i-\bar{X}_{1T})^2/M_4$$

$$\sigma^2_{2T} = \sum_{i=M_4+1}^{M_3}(Z_i-\bar{X}_{2T})^2/(M_3-M_4)$$

$$\sigma^2_{NT} = \sum_{i=M_3+1}^{M_1}(Z_i-\bar{X}_{NT})^2/(M_1-M_3)$$

As the variance of each of the components is inversely proportional to the sample size, the results of the preceding section can be applied to determine the optimal values of the design parameters m_1, m_2 and λ. Of course in most practical situations the optimal design parameters can only be estimated as the population parameters given above are unknown and must be estimated.

AN APPLICATION TO AN NCHS DATA SYSTEM

Estimates of the various population parameters which are required to estimate the optimal values of the design parameters in the example above are not usually published and it is rather difficult to construct realistic "real world" examples. However, using data from Thornberry and Massey (1978), the Statistical Abstract of the United States 1977, published data from the Health Interview Survey (HIS) of NCHS, and making a few fairly reasonable assumptions it is possible to apply the results of the preceding section to HIS.

First the simplifying assumptions are

(1) $\bar{X}_{1T} = \bar{X}_{2T}$

(2) $\sigma^2_{1T} = \sigma^2_{2T}$, and

(3) approximately five percent of telephone households have more than one telephone

Then using the cited data sources together we have

$$M_1 = 72.9 \text{ X } 10^6$$
$$M_2 = 69.3 \text{ X } 10^6$$
$$M_3 = 66.0 \text{ X } 10^6$$
$$M_4 = 62.7 \text{ X } 10^6$$

and for the variable "number of physician visits per year"

$$\bar{X}_{NT} = 11.0$$

$$\bar{X}_{1T} = \bar{X}_{2T} = 14.2$$

$$\sigma^2_{NT} = 55.0$$

$$\sigma^2_{1T} = {}^2_{2T} = 61.7$$

Using the results of the preceding section together with the estimates of the population parameters given above and assuming that a household interview costs \$100, a telephone interview costs \$50 and the total survey budget is \$4,000,000, we have the following estimates for various design strategies:

Estimator	Sample Size Frame 1	Sample Size Frame 2	Variance	Bias	MSE
Household only	40,000	0	$8.23\text{X}10^{12}$	0	$8.23\text{X}10^{12}$
Telephone only	0	80,000	$4.45\text{X}10^{12}$	$14.46\text{X}10^{6}$	$212\text{X}10^{12}$
X_S	25,347	29,305	$6.75\text{X}10^{12}$	0	$6.75\text{X}10^{12}$
X_G (λ)	28,577	28,577	$6.82\text{X}10^{12}$	0	$6.82\text{X}10^{12}$

The estimator based on households only is given by $X_{G1} + X_{G12}$ and can be considered as roughly equivalent to the usual HIS estimator. The estimator based on telephones only is given by

$$\left(\frac{M_3 N_T + (M_1 - M_3) \bar{N}_{NT}}{M_3 \bar{N}_T} \right) X_{G21}$$

where $\bar{N}_T$ and $\bar{N}_{NT}$ represent the average number of persons per household in telephone and non-telephone households respectively. This estimator inflates the estimate for the telephone population up to the total population level. This is a simplified version of the ratio adjusted estimator suggested by Thornberry and Massey (1978). However, it should be noted that for this particular example the bias of the simplified estimator is no larger than the more complex estimators suggested by Thornberry and Massey which inflate differentially by various demographic subclasses. The results in this table speak for themselves. The two estimators X_S and X_G (λ) are for all intents equivalent with respect to MSE. For the specified survey cost either of these two estimators will yield a MSE approximately 20 percent smaller than the usual HIS estimator. The estimator using telephone data only has a MSE nearly 26 times larger than the HIS estimator. This huge increase in MSE is almost totally due to the bias term which does not decrease with increasing sample size. In fact any of the ratio adjustment procedures for reducing bias proposed by Thornberry and Massey (1978) will still result in estimators whose bias term does not decrease with sample size, so in general the bias term in telephone surveys due to undercoverage dominate the MSE for "large" samples.

REFERENCES

Hartley, H.O., "Multiple Frame Surveys," Proceedings of the American Statistical Association, Social Statistics Section (1964) pp. 203-206.

Hartley, H.O., "Multiple Frame Methodology and Selected Application," Sankhya (1974) pp. 99-118.

Sirken, M.G., "Stratified Sample Surveys with Multiplicity," Journal of the American Statistical Association (1972) pp. 224-227.

Sirken, M.G., "The Counting Rule Strategy in Sample Surveys," Proceedings of the American Statistical Association, Social Statistics Section (1974) pp. 119-123.

Thornberry, O.T., and J.T. Massey, "Correcting for Undercoverage Bias in Random Digit Dialed National Health Surveys," Proceedings of the American Statistical Association, Survey Research Section (1978) pp. 224-229.