HYBRID NETWORK ESTIMATORS

Monroe G. Sirken and Gad Nathan[1]
National Center for Health Statistics

KEY WORDS: Multiplicity surveys; counting rules; rare events.

INTRODUCTION

Given a population containing N persons $I = \{I_1,\ldots,I_\alpha,\ldots,I_N\}$, we assume that a sampling frame, $H = \{H_1,\ldots,H_i,\ldots,H_L\}$, lists L households where the persons are potentially eligible to be enumerated. A counting rule is adopted which specifies the conditions that make a person, I_α, eligible to be enumerated at one or more households. Denote the links between the persons and the households at which they are eligible to be enumerated by the indicator variable:

$$\delta_{\alpha,i} = \begin{cases} 1;\ \text{if } I_\alpha \text{ is linked to } H_i \\ 0;\ \text{otherwise.} \end{cases}$$

$$(\alpha = 1,\ldots,N;\ i = 1,\ldots,L)$$

A simple random sample of l households H_i $(i = 1,\ldots,l)$ is selected and the I_α's linked to these sample units are enumerated in the survey. The network estimator of N is:

$$\hat{N}_M = \frac{L}{l} \sum_{i=1}^{l} \lambda_i \ , \qquad (1)$$

where $\lambda_i = \sum_{\alpha=1}^{N} W_{\alpha i}\ \delta_{\alpha i}$ = the weighted number of persons eligible to be enumerated at H_i $(i = 1,\ldots,l)$. The network estimator is unbiased if and only if:

$$\sum_{i=1}^{L} W_{\alpha i}\ \delta_{\alpha i} = 1,\ (\alpha = 1,\ldots,N).$$

The condition would be satisfied if $s_\alpha \geq 1$ and we let the network weights $W_{\alpha i} = 1/s_\alpha$ $(\alpha = 1,\ldots,N)$, where: $s_\alpha = \sum_{i=1}^{L} \delta_{\alpha i}$ = the number of households at which I_α is eligible to be enumerated in the network survey.

The $W_{\alpha i}$ for the I_α enumerated at the sample household H_i, $(i=1,\ldots,l)$ are based on ancillary information that is collected in the network survey from the households at which I_α was enumerated. For example, if $W_{\alpha i} = 1/s_\alpha$ $(\alpha=1,\ldots,N;\ i=1,\ldots,L)$, ancillary information would be needed to determine the value of s_α for every I_α enumerated at a sample H_i $(i=1,\ldots,l)$. Thus every time I_α is enumerated at a sample household, the household would report the number of other households in the sampling frame H where I_α was enumerable.

The variance of $\hat{N}_M$ is:

$$V(\hat{N}_M) = \Theta\Big[\ \frac{1}{L} \sum_{\alpha=1}^{N} \frac{1-s_\alpha}{s_\alpha} + \sum_{\alpha>\beta}^{N} \frac{2}{L} \sum_{i=1}^{L} \frac{\delta_{\alpha i}\ \ \delta_{\beta i}}{s_\alpha\ \ s_\beta} + \lambda(1-\lambda)\ \Big], \qquad (2)$$

where $\Theta = (L^2/l)[(L-l)/(L-1)]$ and $\lambda = N/L$. Assuming that no more than one person is linked to the same household, often a tenable assumption when λ is quite small, (2) reduces to:

$$V(\hat{N}_M) = \Theta[\lambda(h-\lambda)], \qquad (3)$$

where $h = (1/N) \sum_{\alpha=1}^{N} (1/s_\alpha)$ = the inverse of the harmonic mean of the s_α's.

It is of interest to note that $\hat{N}_C$, the estimator of a conventional survey (a survey in which every I_α is uniquely linked to one and only one household, say by the de-jure rule), is a special case of (1) in which $s_\alpha = 1$, $(\alpha=1,\ldots,N)$ and h=1. Substituting h=1 in (3), we obtain:

$$V(\hat{N}_C) = \Theta\ [\lambda\ (1-\lambda)], \qquad (4)$$

assuming again that $\sum_{\alpha=1}^{N} \delta_{\alpha i} \leq 1$. Since $h \leq 1$ in the network survey it follows that:

$V(\hat{N}_C)-V(\hat{N}_M) = \Theta\lambda(1-h) \geq 0$. On the other hand, the conventional estimator is more economical since by design $W_{\alpha i} = 1$, $(\alpha=1,\ldots,N;\ i=1,\ldots,L)$ and ancillary information is not required to calculate the network weights.

HYBRID ESTIMATORS

We have noted that the network estimator $\hat{N}_M$ requires the $W_{\alpha i}$ for every I_α that is enumerated at a sample household. It may happen, however, that not all the households at which I_α is enumerable are able to provide the ancillary information needed to determine $W_{\alpha i}$. We now consider an alternative estimator which counts I_α $(\alpha = 1,\ldots,N)$ at every household where the person is enumerated in the network survey but which requires the $W_{\alpha i}$ for one and only one of these households, which we shall refer to as the key household and could be the de-jure household.

[1] On sabbatical from Hebrew University, Jerusalem.

Similarly to Casady, Nathan and Sirken (1985), we specify that:

$$\delta_{\alpha i} = \delta'_{\alpha i} + \delta''_{\alpha i}, \tag{5}$$

where

$$\delta'_{\alpha i} = \begin{cases} 1; \text{ if } H_i \text{ is the key household of } I_\alpha \\ 0; \text{ otherwise} \end{cases}$$

and

$$\delta''_{\alpha i} = \begin{cases} 1; \text{ if } H_i \text{ is not the key household of } I_\alpha, \text{ but } I_\alpha \text{ is enumerable at } H_i \\ 0; \text{ otherwise.} \end{cases}$$

In other words, the survey adopts a counting rule which specifies two conditions for linking persons to households. For instance, the first condition might be a conventional counting rule, such as the de-jure residence rule, which makes each I_α (α=1,...,N) enumerable at one and only one household, and the second condition makes some or all of the I_α's enumerable at additional households. In the survey, I_α (α = 1,...N) would be enumerated at every eligible household to which he is linked by either condition of the counting rule but the ancillary information would be collected only from households that are linked to persons by the conventional rule.

The hybrid estimator is:

$$\hat{N}_H = \frac{\hat{S}_M}{\hat{S}_C} \hat{N}_C, \tag{6}$$

where: $\hat{N}_C$ is the conventional estimator of N,

$$\hat{S}_M = (L/l) \sum_{\alpha=1}^{N} \sum_{i=1}^{l} \delta_{\alpha i}$$

and $\hat{S}_C = (L/l) \sum_{\alpha=1}^{N} \sum_{i=1}^{l} s_\alpha \; \delta'_{\alpha i}$.

$\hat{N}_H$ is a consistent estimator of N, since:

$$E(\hat{S}_M) = E(\hat{S}_C) = \sum_{\alpha=1}^{N} \sum_{i=1}^{L} \delta_{\alpha i} = S$$

= the number of links between persons and households.

Interest in the hybrid estimator was stimulated by a national household sample survey of heroin users which adopted a counting rule that linked heroin users (1) to their de-jure households and (2) to the households of their "best" friends. The first condition assumes that every heroin user has a de-jure residence. The second condition was added to improve the precision of the survey estimate of N, the number of heroin users. Although the survey encountered little difficulty in enumerating heroin users at the homes of their best friends, it was often unsuccessful in obtaining from them the ancillary information needed to calculate the network weights because the persons who reported a best friend as a heroin user were sometimes uncertain how many other persons that were best friends of the user knew that he was a heroin user. However it was thought that the heroin user could give the information required on the number of persons who regard him as their best friend.

The large sample approximation of the variance of $\hat{N}_H$ is:

$$V(\hat{N}_H) = N^2 \left[\frac{V(\hat{S}_M)}{S^2} + \frac{V(\hat{S}_C)}{S^2} + \frac{V(\hat{N}_C)}{N^2} + \frac{2Cov(\hat{S}_M,\hat{N}_C)}{SN} - \frac{2Cov(\hat{S}_M,\hat{S}_C)}{S^2} - \frac{2Cov(\hat{S}_C,\hat{N}_C)}{SN} \right] \tag{7}$$

It is easily verified that:

$V(\hat{S}_M) = Cov(\hat{S}_M,\hat{S}_C) = \bar{S}\; Cov\; (\hat{S}_M,\hat{N}_C) = \theta\lambda\bar{S}(1-\lambda\bar{S})$;

$Cov(\hat{S}_C,\hat{N}_C) = \theta\lambda\bar{S}(1-\lambda)$;

$V(\hat{S}_C) = \theta\lambda[V(s_\alpha) + \bar{S}^2(1-\lambda)]$,

where $\bar{S} = S/N$ and $V(s_\alpha) = (1/N) \sum_{\alpha=1}^{N} (s_\alpha - \bar{S})^2$;

and by (4), we have $V(\hat{N}_C) = \theta\lambda(1-\lambda)$.

Substituting these expressions in (7), we obtain:

$$V(\hat{N}_H) = \frac{\theta\lambda}{\bar{S}^2} [\; V(s_\alpha) + \bar{S}(1-\lambda\bar{S}) \;]. \tag{8}$$

If the ancillary information needed for the weights were available from the households of drug users' friends as well as from the households of the users themselves, it would have been possible to use $\hat{N}_M$ to estimate N, the number of heroin drug users. But this information was not ascertained for about 20 percent of the users enumerated at the households of their best friends. The loss of sampling efficiency in using $\hat{N}_H$ instead of $\hat{N}_M$ is obtained by subtracting (3) from (8). After some simplification, we obtain:

$$V(\hat{N}_H) - V(\hat{N}_M) = \theta\lambda \left[\frac{V(s_\alpha)}{\bar{S}^2} - \frac{h\bar{S}-1}{\bar{S}} \right]. \tag{9}$$

It can be shown that the bracketed term on the right side of the equation is nonnegative and it is equal to zero if and only if $V(s_\alpha)$=0.

If the ancillary information was collected only from the households at which the heroin user himself was enumerated, it would be possible to estimate N by either $\hat{N}_H$ or $\hat{N}_C$.

Hence, it is of interest to compare the sampling variances of these estimators. By letting h=1 in (9), we have:

$$V(\hat{N}_H) - V(\hat{N}_C) = \theta\lambda \left[\frac{V(s_\alpha)}{\bar{S}^2} - \frac{\bar{S}-1}{\bar{S}} \right]. \quad (10)$$

If $V(s_\alpha) \leq \bar{S}(\bar{S}-1)$, $V(\hat{N}_H) \leq V(\hat{N}_C)$. Otherwise $V(\hat{N}_H) > V(\hat{N}_C)$.

Estimates of the parameters listed in Table 1 are based on preliminary findings of national household sample survey of heroin use in which heroin users were enumerable at their de-jure households and at the households of their best friends.

Table 1: Estimates of Network Parameters Derived from a Household Sample Survey of Heroin Use

Parameter	Estimate
$\bar{S}$	5.6
$V(s_\alpha)$	8.6
h	0.25
λ	0.035

Based on these estimates, design effects of the estimators, (defined as the ratio of their variance to that of the conventional estimator), were estimated and are shown in table 2.

Table 2: Design Effects of the Specified Estimators of Heroin Use Prevalence

Type of Estimator	Symbol	Design Effect
Conventional	$\hat{N}_C$	1
Network	$\hat{N}_M$	0.223
Hybrid	$\hat{N}_H$	0.433

The variance of the hybrid estimator, $\hat{N}_H$, is almost twice as large as the variance of the network estimator $\hat{N}_M$, but it is less than half as large as the variance of the conventional, estimator, $\hat{N}_C$.

LINEAR COMBINATION ESTIMATOR

Since $\hat{N}_H$ and $\hat{N}_C$ are unbiased estimates of N, the linear combination estimator:

$$\hat{N}^* = a(\hat{N}_C) + (1-a)\hat{N}_H = \hat{N}_H + a(\hat{N}_C - \hat{N}_H) \quad (11)$$

is also unbiased, and its variance is:

$$V(\hat{N}^*) = V(\hat{N}_H) + a^2 V(\hat{N}_C - \hat{N}_H) + 2a\rho[V(\hat{N}_H)V(\hat{N}_C - \hat{N}_H)]^{\frac{1}{2}}, \quad (12)$$

where:

$$\rho = \frac{Cov[\hat{N}_H, (\hat{N}_C - \hat{N}_H)]}{[V(\hat{N}_H)\ V(\hat{N}_C - \hat{N}_H)]^{\frac{1}{2}}}$$

$$= \frac{Cov(\hat{N}_H, \hat{N}_C) - V(\hat{N}_H)}{[V(\hat{N}_H)]^{\frac{1}{2}}\ [V(\hat{N}_H) + V(\hat{N}_C) - 2Cov(\hat{N}_H, \hat{N}_C)]^{\frac{1}{2}}}$$

$$= \frac{-\ V(s_\alpha)}{[V(s_\alpha) + S(\overline{1}-\lambda S\overline{)}]^{\frac{1}{2}}\ [V(s_\alpha) + S(\overline{S}-\overline{1})]^{\frac{1}{2}}}. \quad (13)$$

It is noteworthy that $\hat{N}^*$ requires no more information than is required by the hybrid estimator $\hat{N}_H$. That is, the individual is counted at every eligible household but the network weights, W_α, are required for the key households only.

The value of "a" that minimizes (12) is:

$$a = \frac{Cov[\hat{N}_H, (\hat{N}_C - \hat{N}_H)]}{V(\hat{N}_C - \hat{N}_H)}. \quad (14)$$

Let "a" be defined by (13), so that the variance of $\hat{N}^*$ becomes:

$$V(\hat{N}^*) = V(\hat{N}_H)(1-\rho^2).$$

The factor $1-\rho^2$ measures the maximum reduction in the variance that is attainable if N is estimated by $\hat{N}^*$ instead of by $\hat{N}_H$.

Assuming the values of the network parameters listed in Table 1, ρ = .406. The factor $1-\rho^2$ = .16 represents the reduction in the variance of the number of heroin users if the optimum linear combination estimator were used instead of the hybrid estimator.

REFERENCES

Sirken, M.G. (1970), "Household Surveys with Multiplicity", *Journal of the American Statistical Association*, 65, 257-266.

Casady, R.J., Nathan, G. and Sirken, M.G. (1985), "Alternative Dual System Network Estimators", *International Statistical Review*, 53, 183-197.