A DUAL SYSTEM NETWORK ESTIMATOR

Monroe G. Sirken
National Center for Health Statistics

INTRODUCTION

The unique feature of the dual system estimator is that it adjusts for population under-enumeration due to incomplete reporting. It is based on the estimate derived from each of two imperfect data collection systems and on an estimate of the number of the same individuals that are counted by both imperfect systems. The dual system estimator of N, the population size, is

$$\tilde{N} = \frac{\hat{N}_1 \hat{N}_2}{\hat{N}_{12}}$$

where $\hat{N}_1$ is the estimate of N based on the first data collection system, $\hat{N}_2$ is the estimate based on the second system, and $\hat{N}_{12}$ is the estimate of the number of individuals that are counted by $\hat{N}_1$ and $\hat{N}_2$.

Even though $\hat{N}_1$ and $\hat{N}_2$ are biased estimates of N, the dual system estimator of $\tilde{N}$ is a consistent estimate of N if the following assumptions hold:

(1) Every individual in the population is eligible to be enumerated by one or both data collection systems.
(2) Whether the individual eligible to be enumerated by a system is reported or not reported is a Bernoulli variable.
(3) Whether the individual, eligible to be enumerated by both systems, is reported or not reported by one system is independent of whether or not he is reported by the other system.

The traditional dual system estimator assumes that $\hat{N}_1$ and $\hat{N}_2$ are based on different collection systems, such as a household sample survey and a registration system, and that $\hat{N}_{12}$ is determined by matching the files of the individuals enumerated by the two systems. The dual system network estimator is based on procedures that are quite different.

The dual system network estimator assumes that $\hat{N}_1$ and $\hat{N}_2$ respectively are based on disjoint counting rules r_1 and r_2 adopted collaterally by a household survey henceforth referred to as a dual system survey. $\hat{N}_{12}$ is based on a quality check survey using the following scheme that was proposed by Nathan [2]. Quality check interviews are conducted at households that are eligible to report individuals by a different counting rule than the one by which they were enumerated originally in the dual system survey. Thus, if an individual was reported in compliance with r_1 in the the dual system survey, he would be enumerated at a quality check household eligible to report him in compliance with r_2 and vice versa. The objective of the quality check is to ascertain whether or not the individual previously reported in the dual system survey by one rule is also reported by another household that is eligible to report him by the other rule.

For instance, r_1 could denote the de jure residence rule that counts individuals at their usual places of residence, and r_2 a family network rule that counts individuals at residences of their close relatives. Individuals that had been reported by their usual places of residence in the dual system survey, would be enumerated at the households of close relatives in the quality check survey, and the individuals that had been reported by relatives' households in the dual system survey would be enumerated at their own households in the quality check survey. The quality check households are identified by respondents in the dual system survey. Thus, the individuals that report themselves in the dual system survey provide the names and addresses of relatives that would be eligible to report them in the quality check survey, and the relatives who report individuals in the dual system survey provides the names and addresses of these individuals for follow-up in the quality check survey.

The dual system network estimator and its component estimators are derived in the following sections of this paper. Biases of the traditional dual system estimator and of the dual system network estimator are briefly discussed in the final section.

ESTIMATORS $\hat{N}_1$ AND $\hat{N}_2$

The I_α (α=1,...,N) individuals in a population are enumerated in the dual system household survey by two mutually exclusive counting rules, r_1 and r_2. The links between I_α (α=1,...,N) and the households H_i (i=1,...,M) eligible to report them in compliance with these rules are specified by

$$\delta_{\alpha i} = \delta_{\alpha i1} + \delta_{\alpha i2}$$

where the indicator variable

$$\delta_{\alpha ik} = \begin{cases} 1 & \text{if } I_\alpha \ (\alpha=1,\ldots,N) \text{ is linked to } H_i \\ & (i=1,\ldots,M) \text{ by } r_k \ (k=1,2) \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$N_k = \sum_{\alpha=1}^{N} \sum_{i=1}^{M} \delta_{\alpha ik}/s_{\alpha k} = \text{the number of the}$$

I_α (α=1,...,N) eligible to be enumerated by r_k (k=1,2) and

$$N_{12} = \sum_{\alpha=1}^{N} \sum_{i=1}^{M} \delta_{\alpha i1}/s_{\alpha 1} \sum_{j=1}^{M} \delta_{\alpha j2}/s_{\alpha 2} = \tag{1}$$

the number of the I_α (α=1,...,N) eligible to be enumerated by r_1 and r_2.

where

$$s_{\alpha k} = \sum_{i=1}^{M} \delta_{\alpha ik} = \text{the number of households}$$

eligible to report I_α (α=1,...,N) by r_k (k=1,2). If $s_{\alpha k} \geq 1$ (α=1,...,N) (k=1,2), $N_k = N_{12} = N$. On

the other hand, if $s_{\alpha 1} \geq 1$ $(\alpha=1,\ldots,N)$ but $s_{\alpha 2} = 0$ for some I_α, $N_1 = N$ but $N_2 = N_{12} \neq N$. And if $s_{\alpha 1} = 0$ and $s_{\alpha 2} = 0$ for some I_α, $N_1 \neq N_{12}$ and $N_2 \neq N_{12}$ unless the identical I_α are covered by r_1 and r_2.

The I_α $(\alpha=1,\ldots,N)$ may or may not be reported in the survey by the H_i $(i=1,\ldots,M)$ to whom they are linked by r_1 and r_2. Consequently, we denote the survey responses by the random variables

$$\tilde{\delta}_{\alpha ik} = \begin{cases} 1 & \text{if } I_\alpha \ (\alpha=1,\ldots,N) \text{ is linked to and enumerated at } H_i (i=1,\ldots,M) \text{ by } r_k \ (k=1,2) \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the $E(\tilde{\delta}_{\alpha ik}) = p_k \, \delta_{\alpha ik}$ $(\alpha=1,\ldots,N)$ $(i=1,\ldots,M)$ $(k=1,2)$, where $0 < p_k \leq 1$.

A random sample of m out of M households is selected, and the I_α $(\alpha=1,\ldots,N)$ at the m households are enumerated by r_1 and r_2 in the dual survey. Two estimates of N are derived from the survey.

$$\hat{N}_k = \frac{M}{m} \sum_{i=1}^{M} a_i \sum_{\alpha=1}^{N} \tilde{\delta}_{\alpha ik}/s_{\alpha k} \quad (k=1,2) \tag{2}$$

where the Bernoulli variable

$$a_i = \begin{cases} 1 & \text{if } H_i \ (i=1,\ldots,M) \text{ is selected in the dual system survey} \\ 0 & \text{otherwise.} \end{cases}$$

Since $E(\hat{N}_k) = N_k \, p_k$ $(k=1,2)$ it follows that $\hat{N}_k$ is an unbiased estimate of N if and only if $s_{\alpha k} \geq 1$ $(\alpha=1,\ldots,N)$ and $p_k = 1$.

ESTIMATOR $\hat{N}_{12}$

Subsamples of reports of individuals enumerated by r_1 and r_2 in the dual system survey are selected for the quality check survey. For each type of report, a quality check household is selected at random from $s_{\alpha t}$ households $(t=3-k)$ where the I_α $(\alpha=1,\ldots,N)$ is enumerable by a different rule than the one by which he was enumerated originally. Thus, if I_α had been enumerated by r_k $(k=1,2)$ in the dual system survey, a household eligible to report him by r_t $(t=3-k)$ would be selected for the quality check survey. The quality check survey ascertains whether or not the I_α $(\alpha=1,\ldots,N)$ that were reported by r_k $(k=1,2)$ in the dual system survey will also be reported by households eligible to report them by r_t. Obviously, it would be impossible to select quality check households either for the individuals that were enumerated by r_k $(k=1,2)$ if $s_{\alpha t} = 0$, or for the individuals that were not enumerated by r_k because $s_{\alpha k} = 0$.

The selection of individuals in the quality check survey is represented by the Bernoulli variable

$$\varepsilon_{\alpha ik} = \begin{cases} 1 & \text{if } \tilde{\delta}_{\alpha ik} = 1, \ s_{\alpha t} > 0 \text{ and the dual system survey report of } I_\alpha \ (\alpha=1,\ldots,N) \text{ by } H_i \ (i=1,\ldots,M) \text{ using } r_k \ (k=1,2) \text{ is selected in the quality check survey} \\ 0 & \text{otherwise.} \end{cases}$$

If $s_{\alpha t} > 0$ $(t=3-k)$, it follows that $E(\varepsilon_{\alpha ik} | \tilde{\delta}_{\alpha ik}) = f_k \, \tilde{\delta}_{\alpha ik}$ $(k=1,2)$ where f_k is the rate at which the I_α $(\alpha=1,\ldots,N)$ reported by rule r_k in the dual system survey are selected in the quality check survey. If $s_{\alpha t} = 0$, $E(\varepsilon_{\alpha ik} | \tilde{\delta}_{\alpha ik}) = 0$. Since the individuals selected in the quality check survey may or may not be reported by the quality check households eligible to report them, let

$$\tilde{\varepsilon}_{\alpha ik} = \begin{cases} 1 & \text{if } \varepsilon_{\alpha ik} = 1 \ (\alpha=1,\ldots,N)(i=1,\ldots,M) \ (k=1,2), \text{ and } I_\alpha \text{ is reported by the selected quality check household using } r_t \ (t=3-k) \\ 0 & \text{otherwise.} \end{cases}$$

Assuming that (1) the probability of being enumerated by a particular counting rule is the same in the dual system survey and in the quality check survey, and (2) the chances of being enumerated in both surveys are independent, we have

$$E(\tilde{\varepsilon}_{\alpha ik} | \varepsilon_{\alpha ik}) = p_t \, \varepsilon_{\alpha ik}$$

and

$$E(\tilde{\varepsilon}_{\alpha ik}) = \begin{cases} f_k \, p_k \, p_t \, \delta_{\alpha ik} \ (k=1,2) \text{ if } s_{\alpha t} > 0 \ (t=3-k) \\ 0 \quad \text{otherwise.} \end{cases}$$

Finally, we have two estimators of the N_{12}, previously defined by (1) as the number of the I_α $(\alpha=1,\ldots,N)$ eligible to be enumerated by r_1 and r_2.

$$\hat{N}_{12k} = \frac{M}{m f_k} \sum_{i=1}^{M} \alpha_i \sum_{\alpha=1}^{N} \tilde{\varepsilon}_{\alpha ik}/s_{\alpha k} \quad (k=1,2). \tag{3}$$

Under the assumptions specified in the preceding paragraph,

$$E(\hat{N}_{12k}) = N_{12} \, p_1 \, p_2. \tag{4}$$

ESTIMATOR $\tilde{N}$

The dual system network estimator of N is

$$\tilde{N} = \frac{\hat{N}_1 \hat{N}_2}{\hat{N}} \tag{5}$$

where

$$\hat{N}_{12} = \beta \hat{N}_{121} + (1-\beta) \hat{N}_{122} \quad (0 \leq \beta \leq 1). \tag{6}$$

Assuming large samples, and independence of response in the dual system survey and the quality check survey, we have

$$E(\tilde{N}) \approx \frac{E(N_1) \ E(\tilde{N}_2)}{E(\hat{N}_{12})} = \frac{N_1 N_2}{N_{12}} \tag{7}$$

If $s_{\alpha k} \geq 1$ $(\alpha=1,\ldots,N)(k=1,2)$, $N_1 = N_2 = N_{12} = N$, and it follows that $E(\tilde{N}) = N$. On the other hand,

if $s_{\alpha 1} \geq 1$ (α=1,...,N) but $s_{\alpha 2} = 0$ for some of the I_α, $N_1 = N$ but $N_2 = N_{12} \neq N$. Nevertheless, under these latter conditions it also follows that $E(\hat{N}) = N$. Thus, we have shown that $\hat{N}$ is a consistent estimate of N if $s_{\alpha k} \geq 1$ (α=1,...,N) for either or both counting rules, and if the responses in the dual system survey and the quality check surveys are independent.

CONCLUDING REMARKS

The dual system network estimator has the same form as the traditional dual system estimator. For the traditional estimator, $\hat{N}_1$ and $\hat{N}_2$ are based on the outcomes of two separate data systems, a household survey and registration system respectively, and $\hat{N}_{12}$ is obtained by matching individuals that are reported by both systems. It has been shown that the household survey used by the traditional dual system estimator may be based on either traditional sampling [1] or network sampling [3].

The dual system traditional estimator and the dual system network estimator are subject to correlation bias if the assumption of response independence is not satisfied. The traditional estimator would have correlation bias if the responses to the two data systems were dependent. The network estimator would have correlation bias if responses to the two counting rules were dependent.

The dual system estimators are also subject to other types of nonsampling errors. For instance, the traditional estimator is subject to matching bias if individuals reported by both data systems are not matched or if individuals not reported by both systems are erroneously matched. Matching bias is likely to occur if the variables used for matching are based on poor quality data. The network estimator is subject to address bias if the addresses of the quality check households can not be located. Address bias occurs when the dual survey households provide incomplete or inaccurate addresses for the quality check households [4].

Overall, however, relatively little is known about the error effects of the traditional dual system estimator and even less is known about the error effects of the dual system network estimator.

ACKNOWLEDGEMENT

The author thanks Gad Nathan for his many helpful comments.

REFERENCES

[1] Marks, Eli S., William Selzer, and Karol J. Krotki, Population Growth Estimation, The Population Council, New York, 1974.

[2] Nathan, Gad, "An Empirical Study of Response and Sampling Errors for Multiplicity with Different Counting Rules," JASA, Vol. 71, No. 356, Dec. 1976, pp. 808-15.

[3] Sirken, Monroe G., "Dual System Estimators Based on Multiplicity Surveys," in Developments in Dual System Estimation of Population Size and Growth, the University of Alberta Press, 1978, pp. 81-88.

[4] Sirken, Monroe G., Barry I. Graubard and Richard W. LaValley, "Evaluation of Census Population Coverage by Network Surveys," 1978 Proceedings of the Section on Survey Research Methods, pp. 239-244.