ESTABLISHMENT SURVEY DESIGNS: CONVENTIONAL SAMPLING VERSES NETWORK SAMPLING

Monroe G. Sirken
National Center For Health Statistics

A. Introduction

This paper is about the effect of sampling frames on the variances of establishment survey estimators of the quantity X, where X is the sum of the x-variate over the M transactions that R establishments have with N households.

Let M_{ij} = the number of transactions between the E_j (j = 1, 2, ..., R) establishment and the H_i (i = 1, 2, ..., N) household. Then $M_j = \sum_{i=1}^{N} M_i$ = the number of transactions of E_j with N households, and $M = \sum_{j=1}^{R} M_j$. Also, let X_{jk} denote the value of the x-variate for the k (k = 1, ..., M_j) transaction of the E_j establishment. Then $X_j = \sum_{k=1}^{M_j} X_{jk}$ = the sum of the x-variate over the M_j transactions of E_j , and $X = \sum_{j=1}^{R} X_j$.

Establishment surveys based on conventional and network sampling differ with respect to the kinds of listings used for the establishment sampling frame. Conventional establishment surveys use free-standing sampling frames, and network sampling establishment surveys use population survey-generated sampling frames.

The free-standing sampling frame is a complete and unduplicated listing of R establishments, E_j (j=1,2,...,R), and their respective measures of size, M_j (j = 1, 2, ..., R).

The population survey-generated frame is a listing of n sample households H_i (i = 1, 2, ..., n) that were selected in a household sample survey, and numbers of transactions, M_{ij},that each sample household reported having with each distinct establishment E_j (j = 1 , 2 , ..., R).

The free-standing frame is compiled from one or more independent sources of information, such as membership listings of professional and commercial associations, telephone and other directories, etc. The survey generated frame is based on information reported by households enumerated in a population sample survey. Household respondents identify establishments with whom they have transactions, and they report the numbers of their transactions with each establishment.. It is noteworthy that survey respondents do not report x-variates for their own transactions or for transactions of other households.

In the free-standing frame, every listing unit represents a distinct establishment. Thus, there is a one-to-one correspondence between R listing units and R establishments. In the survey-generated frame, n sample households are the listing units, and listing units and establishments are not in one-to-one correspondence. Each of the n households is linked to a cluster of between 0 - R establishments with whom it has transactions, and the same establishments may be linked to multiple listing units because each establishment is linked to a network of between 1 - N households with whom it has transactions. Conventional sampling applies in establishment surveys using free-standing frames because every establishment is uniquely linked to one and only one listing unit, and network sampling applies in establishment surveys using survey-generated frames because the same establishments may be linked to multiple listing units (Sirken, 1998).

B. Conventional Sampling Establishment Survey

The conventional establishment survey is a two-stage self-weighted probability- proportionate-to-size (PPS) sample survey in which single establishments are first stage selection units, and their transactions are second stage selection units. Assuming a sample of r establishments is selected with probability proportionate-to-size and with replacement, and fixed subsamples of c transactions each © is a positive integer greater than zero) is selected by simple random sampling without replacement per selected first stage unit, the unbiased conventional establishment survey estimator of X is

$$X_C' = \frac{M}{rc} \sum_{j=1}^{r} X_j' \qquad (1)$$

where X_j is the sum of the x-variate over the sample of c transactions linked to first stage selection unit j.

The variance of X_C is (Thompson, 1992)

$$\text{Var}(X_C) = \frac{M^2}{r}\sigma^2_{CB} + \frac{M}{cr}\sum_{j=1}^{R}(M_j - c)\sigma_j^2 \qquad (2)$$

where the first term on the right side of (2) is the first stage variance component of X_C and

$$\sigma^2_{CB} = \frac{1}{M}\sum_{j=1}^{R} M_j(\bar{X}_j - X/M)^2 \qquad (3)$$

is the between establishment component of total population variance. The second term on the right side of (2) is the second stage variance component of X_C, and

$$\sigma_j^2 = \frac{1}{M_j - 1}\sum_{k=1}^{c}(X_{jk} - X_j/M_j)^2 \qquad (4)$$

is the within establishment population variance of the E_j establishment.

C. Network Sampling Establishment Survey

The establishment survey using network sampling is a two-stage sample survey in which households are first stage selection units and transactions of establishments linked to first stage units are second stage units. The network establishment survey is self-weighted if the population survey generating its sampling frame is self-weighted, and it is PPS if $N_j/N = M_j/M$, where N_j = the number of households having transactions with the E_j (j = 1, 2, ..., N) establishment.

Assuming the n households listed in the survey-generated frame are first stage selection units are selected by s r s with replacement, and that a sample of s times M_{ij} (s is a positive integer greater than zero) transactions is selected by s r s without replacement every time the E_j (j = 1, 2, ..., R) establishment is linked to a sample household H_i (i = 1, 2, ..., n), the unbiased self weighting network sampling establishment survey estimator of X is

$$X_N = \frac{N}{n}\sum_{i=1}^{n}\sum_{j\in A_i} X_j(i) \qquad (5)$$

where A_i is the cluster of establishments that is linked to selection unit i , and $X_j(i)$ is the sum of the variate over the sample of the sM_{ij} transactions of the E_j establishment belonging to A_i. Under the conditions specified X_N is not necessarily a PPS estimator.

The variance of X_N is (Sirken, Shimizu, Judkins, 1998)

$$\text{Var}(X_N) = \frac{N^2}{n}\sigma^2_{NB} + \frac{N}{sn}\sum_{i=1}^{N}\sum_{j=1}^{R} M_{ij}\frac{M_j - sM_{ij}}{M_j}\sigma_j^2 \qquad (6)$$

where the first term on the right side of (6) is the first stage component of variance of the network sampling estimator, and

$$\sigma^2_{NB} = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j\in A_i} M_{ij}\bar{X}_j - X/N)^2 \qquad (7)$$

is the between household component of the population variance, and the second term on the right side of (7) is the second stage variance component of the network estimator.

D. Difference Between Variances In Conventional and Network Sampling

Let m_C = r c = the transaction sample size in the conventional sampling establishment survey, where r is the number of first stage establishment selections , and c is the number of second stage transactions selected per first stage selection unit. And let $m_N = sm_N^*$ = the transaction sample size in the network sampling establishment survey, where s is the number of transactions selected per transaction linked to n first stage sample households, and

$$m_N^* = \sum_{i=1}^{n}\sum_{j\in A_i} M_{ij} = \text{sum of the transactions}$$

that are linked to n sample households.

Clearly, m_N^* is a random variable and its expected value, conditional on a sample of n households, is

$$E(m_N^*|n) = n\varphi$$

where

$$\varphi = M/N = \text{number of transactions per household.}$$

In the comparisons of the conventional and network variances that follows, the sample sizes of the two surveys are equated by letting $r = E(m_N^*|n) = n\varphi$ and c = s. Thus, it follows that the expected sample sizes of the transactions selected are the same in both surveys, that is, $E(m_C) = E(rc) = sE(m_N^*|n) = sn\varphi = E(m_N)$. Calibrating the sample sizes in this manner, not only assures that the total transaction sample sizes are the same in both surveys but also enhances the prospects of selecting very roughly about the same number of distinct establishments in both surveys. Thus, the calibration serves as a very rough approximation to conducting both surveys under the same cost constraints because total data collection costs in two-stage surveys are essentially cost functions of the numbers of distinct first and second stage selection units.

After substituting r = $n\varphi$ and c = s in the formula for the conventional sampling and remembering that $M = N\varphi$, in (2), the variance of X_C becomes

$$\text{Var}(X_C') \approx \frac{N^2\varphi}{n}\sigma_{CB}^2 + \frac{N}{sn}\sum_{j=1}^{R}\frac{M_j - s}{M_j - 1}M_j\sigma_j^2, \quad (8)$$

and the difference between the network and conventional sampling variances shown in (6) and (8) respectively reduces to

$$Var(X_N') - Var(X_C') = \frac{N^2}{n}[\sigma_{NB}^2 - \varphi\sigma_{CB}^2]$$

$$-\frac{N}{sn}\sum_{j=1}^{R}\frac{\sigma_j^2}{M_j}\sum_{i=1}^{N}M_{ij}(M_{ij}-1). \quad (9)$$

The first term on the right side of (9) is the difference between the first stage variance components; it represents the difference between the between-household variance component in network sampling and the between-establishment variance component in conventional sampling. The second term on the right side of (9) is the difference between the second stage variance components; it represents the conventional and network sampling difference in the within-establishment variance component.

First stage variance difference

It is apparent from (9) that first stage variances in network and conventional sampling are equal if

$\sigma_{NB}^2 = \varphi\sigma_{CB}^2$, and the first stage variance is less for network sampling than conventional sampling if .

$\sigma_{NB}^2 < \varphi\sigma_{CB}^2$, and it is greater for network than for

conventional sampling if $\sigma_{NB}^2 > \varphi\sigma_{CB}^2$.

If the following conditions exist in which

$\varphi = M/N \le 1$, then $\sigma_{NB}^2 - \varphi\sigma_{CB}^2 \ge \sigma_\psi^2$ where ψ is a discrete random variable with the probability distribution

$$Pr(\psi = X/N) = 1 - \varphi$$

and

$$Pr(\psi = X/N - X/M) = \varphi$$

and

$$\sigma_\psi^2 = \varphi(1-\varphi)(X/M)^2.$$

The proof of each of each of the following statements appears in the appendix..

Condition 1. If none of the N household has multiple transactions, then φ = M/N = the fraction of households each having a single transaction, and $1 - \varphi$ = the fraction of households without any transactions, then

$$\sigma_{NB}^2 - \varphi\sigma_{CB}^2 = \sigma_\psi^2$$

For this condition, $\sigma_{NB}^2 = \sigma_{CB}^2$ if $\varphi = 1$, and $\sigma_{NB}^2 = \sigma_{CB}^2 = 0$ if $\varphi = 0$.

Condition 2. If one or more of the N households have multiple transactions and

$\bar{X}_j \le 0$ (j = 1, 2, ... , R), then

$$\sigma_{NB}^2 - \varphi\sigma_{CB}^2 > \sigma_\psi^2.$$

Second stage variance difference

It is apparent from the second term on the right side of (9) that second stage variances of the conventional and network sampling surveys are equal, when

$$\sum_{i=1}^{N}M_{ij}(M_{ij}-1) = 0 \quad (j = 1, 2, ..., R).$$

These conditions are satisfied when none of the N households has multiple transactions with the same establishment and /or if $\sigma_j^2 = 0$ (j = 1, 2, ..., R). When $\sum_{i=1}^{N}M_{ij}(M_{ij}-1) > 0$ for any j, the second stage variance is less for the network sampling than conventional sampling, and the magnitude of the variance difference depends jointly on the extent to which households have multiple transactions with the same establishments and on the σ_j^2's.

Combined first and second stage variance difference

The difference between the sum of the first and second stage variance is necessarily less for conventional than for network sampling whenever $\varphi < 1$ and none of N households has multiple transactions with the same establishments. If $\varphi < 1$ and any of the N households have multiple transactions with the same establishments or if $\varphi > 1$, the direction and magnitude of difference between the combined first and second stage variances may favor either network sampling or conventional sampling depending on the distributions of transactions among the N households and among the R establishments. Multiple transactions with the same establishments would tend to favor network sampling, and multiple transactions with different establishments would tend to favor conventional sampling.

E. Concluding Remarks

This paper proposes a network sampling two-stage establishment survey design to estimate the quantity X, the sum of the x-variate over all transactions

that households have with establishments. The network sample design is not proposed as a substitute for the conventional sampling two-stage PPS establishment survey design under all survey conditions, but as an alternative design especially worthy of serious considerations when (1) stand-alone establishment frames with good measures of size are difficult or impossible to construct or maintain, and (2) population survey-generated sampling frames can be constructed and maintained as adjuncts to on-going household sample surveys.

These findings ignore nonsampling errors and suppose that the stand-alone and population survey-generated establishment frames are available and are in flawless condition. Under these conditions , and supposing that expected sizes of the transaction samples are the same in both surveys, and the expected sizes of the non distinct establishment samples are the same in both surveys, a two-stage sampling error model compares sampling variances of the network and the conventional establishment survey estimators of X.

In most instances, it seems likely that the first stage variance component would be smaller for conventional than network sampling because first stage selection units are single establishments in conventional sampling and they are clusters of varying sizes ranging between 0 to R establishments in network sampling. And in most commonly encountered sample survey designs, sampling variances are generally smaller when selection units are single observation units than when they are clusters containing variable numbers of observation units. On the other hand, second stage variance components would be nearly always smaller for network than conventional sampling if any of the households have multiple transactions with the same establishments, and second stage units are selected without replacement. .

Though the findings in this paper are neither definitive nor conclusive, they are encouraging in indicating that sampling variances are not necessarily larger for network sampling than for conventional sampling, and they are helpful in pinpointing the important areas for further research. For example, with respect to sampling errors, data are needed to compare σ^2_{BN} and σ^2_{BC}, the first stage population variance components in network and conventional sampling respectively. With respect to non sampling errors, information is needed to assess the relative quality and costs of constructing and maintaining population survey-generated sampling frames and free-standing sampling frames.

Appendix: Proofs

Proof of the first condition

If $\varphi = M/N \le 1$ and none of the N households has multiple transactions,

$$\sigma^2_{PB} - \varphi\,\sigma^2_{CB} = \sigma^2_{\psi}$$

Proof of this statement is equivalent to proving that

$$\sigma^2_{NB} = \sigma^2_{NB\cdot} = \varphi\,\sigma^2_{CB} + \sigma^2_{\psi}$$

where $\sigma^2_{NB\cdot}$ represents the between household population variance and because none of the N households have multiple transactions, φ = the fraction of N households that each have one transaction and $1 - \varphi$ = the fraction one of N households withour any transactions..

Because none of the N households have multiple transactions, $M_{ij} = M^*_{ij}$ is a binomial variable, where $M^*_{ij} = 1$ if the H_i (1 = 1, 2, . N) household has a transaction with the E_j (j = 1, 2 , ..., R) establishment, and $M^*_{ij} = 0$ otherwise. Hence,

$\sum_{i=1}^{N} M^*_{ij} = M_j = N_j$ = the number of distinct households having transactions with E_j, and $\sum_{j=1}^{R} N_j = M$. Thus, it follows that

$$\begin{aligned}
\sigma^2_{PB\cdot} &= \frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{j\in A_i} M^*_{ij}\bar{X}_j - X/N\Big)^2 \\
&= \frac{\varphi}{M}\sum_{j=1}^{R} M_j\Big(\bar{X}_j - \frac{X}{N}\Big)^2 \\
&\quad + (1-\varphi)(X/N)^2 \\
&= \varphi\,\sigma^2_{CB} + \varphi(X/M - X/N)^2 \\
&\quad + (1-\varphi)(X/N)^2 \\
&= \varphi\,\sigma^2_{CB} + \sigma^2_{\psi}
\end{aligned}$$

where as shown in (3)

$$\sigma^2_{CB} = \frac{1}{M}\sum_{j=1}^{R} M_j(\bar{X}_j - X/M)^2,$$

and ψ is a discrete random variable with the probability distribution

$$Pr(\psi = X/N) = 1 - \varphi.$$

Thus,

$$Pr(\psi = X/N - X/M) = \varphi$$

so that

$$\sigma_{\psi}^2 = E(\psi^2) = \varphi(X/M - X/N)^2 + (1-\varphi)(X/N)^2 = \varphi(1-\varphi)(X/M)^2.$$

Proof of the second condition

If $\varphi \le 1$, and $X_j \ge 0$ (j = 1, 2, ... , R), and any of the N households have multiple transactions,

$$\sigma_{PB}^2 - \varphi\sigma_{CB}^2 \ge \sigma_{PB*}^2 - \varphi\sigma_{cB}^2 = \sigma_{\psi}^2.$$

Subtracting and adding σ_{PB*}^2 to the term on the left side of the above equation

$$\sigma_{PB}^2 - \varphi\sigma_{CB}^2 = [\sigma_{PB}^2 - \sigma_{PB*}^2] + [\sigma_{PB*}^2 - \varphi\sigma_{CB}^2] = \sigma_{PB}^2 - \sigma_{PB*}^2 + \sigma_{\psi}^2.$$

Supposing $\bar{X}_j \ge 0$ (j = 1, 2, ..., R),

$$\sigma_{PB}^2 - \sigma_{PB*}^2 = \frac{1}{N}\sum_{i=1}^{n}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_j\right)^2 - \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{R} M_{ij}^*\bar{X}_j\right)^2$$

$$\ge \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{R}(M_{ij}\bar{X}_j)^2 - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{R}(M_{ij}^*\bar{X}_j)^2$$

$$= \frac{1}{N}\sum_{j=1}^{R}(\bar{X}_j)^2\sum_{i=1}^{N}(M_{ij}^2 - M_{ij}^*) \ge 0.$$

Thus, it follows that

$$\sigma_{PB}^2 - \varphi\sigma_{CB}^2 \ge \sigma_{\psi}^2.$$

References

Sirken, Monroe G.(1997). Network sampling. *Encyclopedia of Biostatistics* . Wiley and Sons. Vol. 4, 2977-2986.

Sirken, M., Shimizu, I., and Judkins, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1, 470-473.

Thompson, S. (1992). *Sampling* . *New York:* John Wiley and Sons, Inc.