DESIGN EFFECTS OF LINKED POPULATION/ESTABLISHMENT SURVEYS

Monroe G. Sirken and Iris Shimizu
National Center for Health Statistics, 3311 Toledo Road, Room 5212, Hyattsville, Maryland 20782

KEYWORDS: Establishment transactions, Population survey-generated establishment sampling frame; Network sampling, Multiplicity estimator, Conventional survey estimator

A. Introduction[1]

Survey statistics on the number and kinds of transactions that populations have with establishments are extensively used in monitoring and planning the nation's economic, health, and social programs. These kinds of statistics are typically collected by population sample surveys (*ps*) in which households report the variables of interest for their transactions with establishments, or by establishment sample surveys (*es*) in which establishments report the variables of interest about their transactions with populations. The linked population/establishment sample survey (*ls*), though rarely used, is a third survey design option for estimating the volume of transactions between establishments and populations. The *ls* is a potential design alternative to the *es* especially when the *es* stand-alone frame has poor coverage or inadequate size measures. The *ls* is a potential design alternative to the *ps* especially when the transactions of interest refer to rare or elusive populations that are hard to find in the *ps*, or the variables of interest about transactions relate to sensitive or technical topics that are hard to enumerate in the *ps*.

The *ls* idea was proposed more than a decade ago by a Panel of the Committee on National Statistics (CNSTAT) when reviewing plans of the National Center for Health Statistics (NCHS) to restructure the designs of its national health care provider surveys (Wunderlich, 1992). The Panel recommended that the NCHS investigate the possibilities of using the listings of health care providers visited by households in the National Health Interview Survey as sampling frames of its establishment surveys in much the same way that NHIS household rosters were being used as sampling frames of population surveys. This paper summarizes research findings that compare the *ls* sampling errors with the *es* and *ps* sampling errors, and explore the kinds of configurations of transactions between households and establishments that favor the *ls* or one of the other surveys. Notation used in this paper is shown in Exhibit 1.

[1] The opinions expressed in this paper are those of the authors and not necessarily those of the National Center for Health Statistics

B. The unbiased *ls* estimator and its variance

The *ls* is a two-phase hybrid of the *ps* and the *es*. Phase I involves a population sample survey in which households report their transactions and identify the establishments with whom they have transactions but they do not report the variables of interest about their transactions. Phase II involves a follow-up establishment survey with the establishments generated in Phase I and these establishments report the variables of interest for their transactions with all households.

Let X = the x-variable summed over the M transactions of R establishments and N households. The *ls* estimator of X is modeled as the network sampling estimator of a two-stage population sample survey in which households are the primary sampling units and transactions of the establishments with whom the survey households have transactions are second stage sampling units. (A network sampling rather than

Exhibit 1: Notation

Let

M_{ij} = the number of transactions household i (i = 1, 2, …, N) has with establishment j (j = 1, 2, …, R),

$M_{\_j} = \sum_i^N M_{ij}$ = number of transactions establishment j has with all households,

$M = \sum_j^R M_{\_j} = \sum_i^N \sum_j^R M_{ij}$,

$\bar{M} = M/N$ = the average number of transactions per household,

X_{ij} = the sum of the x-variable over the M_{ij} transactions of household I with establishment j (Note that $X_{ijh} = 0$ and, hence, $X_{ij} = 0$ if $M_{ij} = 0$),

$X_{\_jk}$ = the value of the x-variate for transaction k of establishment j with households,

$X_{\_j} = \sum_k^{M_j} X_{\_jk}$ = sum of the x-variate over all transactions with households for establishment j,

$\bar{X}_{\_j} = X_{ij}/M_{\_j}$ = the mean of the x-variate for transactions of establishment j,

$X = \sum_j^R X_{\_j} = \sum_i^N \sum_j^R X_{ij}$,

c_{ls} (>0) is the number of transactions selected in ls (from all of establishment j's transactions) for each transaction which surveyed household i reported having with establishment j,

n = the number of surveyed households,

m = the number of sample transactions, and

r = the number of sample establishments.

a conventional sampling estimator of X is used because transactions of the same establishments are counted at every household with whom the establishments have transactions.) In the first stage, a sample of n households is selected by simple random sampling (srs) with replacement (WR). In the second stage, a sample of the transactions of each establishment linked to a surveyed household is independently selected by srs without replacement (WOR) with transaction sample sizes proportional to the number the establishment's transactions with that household.

The *ls* estimator and variance are shown in Exhibit 2. For every transaction of survey household i (i =1, …, n) with establishment j (j = 1, …, R), the single-stage *ls* estimator counts the parameter $\bar{X}_{\_j} = X_{\_j}/M_{\_j}$, and the two-stage *ls* estimator counts the estimate $\bar{X}'_{\_j} = \left(1/CM_{ij}\right)\sum_k^{CM_{ij}} X_{\_jk}$. The variance of the single-stage *ls* estimator of X depends on the between household population variance, and the variance of the two-stage *ls* estimator of X is the sum of the between and the within household population variance component.

C. Comparing *ls* and *es* sampling errors

The *ls* was originally viewed as an establishment sample survey that uses a population sample survey-generated sampling frame which lists the establishments that have transactions with survey households and the number of their transactions with each household. This section briefly summarizes the research findings which compare the sampling errors of *ls* and *es* estimators of X in two-stage self-weighting sample surveys with equivalent expected transaction sample sizes. The *ls* is designed with a two-stage self-weighted sample in the manner described earlier. The *es* is designed as a two-stage self-weighted sample of transactions in which establishments are selected with probability proportional to transaction size (pps) WR. The *es* estimator and variance are shown in Exhibit 3.

The *ls* and *es* estimators of X are equivalent and their sampling errors are equivalent if, and only if, M transactions of the population that generates the establishment frame are uniformly distributed over N = M households such that each household has one and only one transaction. If the M transactions are not uniformly distributed in this manner, (1) the first stage variance component virtually always favors the *es*, (2) the second stage variance component virtually always favors the *ls* when the *ls* transaction samples are selected are selected WR, and (3) the second stage *ls* and *es* variances are equivalent when the transaction samples are selected WOR. The amount by which the *es* first-stage variance component is favored increases with increases (1) in the fraction of the households without transactions and (2) in the variance of the distribution of the M transactions for the truncated population of households that have transactions. The

Exhibit 2: Estimators and variances for *ls* survey

The **single-stage *ls* estimator** (when transactions are not sampled within establishments) of X is

$$X'_{ls} = \frac{N}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{R} M_{ij}\bar{X}_{\_j}\right] \quad (2.1)$$

with variance

$$Var(X'_{ls}) = \left(N^2/n\right)\sigma^2_{ls1}, \text{ where} \quad (2.2)$$

$$\sigma^2_{ls1} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_{\_j} - \frac{X}{N}\right)^2. \quad (2.3)$$

The **two stage *ls* estimator** (when transactions are sampled within establishments) of X is:

$$X''_{ls} = \frac{N}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{R} M_{ij}\bar{X}'_{\_j}(i)\right], \text{ where} \quad (2.4)$$

$$\bar{X}'_{\_j}(i) = \frac{1}{c_{ls}M_{ij}}\sum_{k}^{c_{ls}M_{ij}} X_{\_jk}$$

is the unbiased estimate of $\bar{X}_{\_j}$. The variance of X''_{ls} is

$$Var(X''_{ls}) = \frac{N^2}{n}\left[\sigma^2_{ls1} + \frac{1}{c_{ls}}\sigma^2_{ls2}\right], \text{ where} \quad (2.5)$$

$$\sigma^2_{ls2} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{R} M_{ij}\left(1 - \frac{c_{ls}M_{ij}}{M_{\_j}}\right)\sigma^2_{\_j} \text{ and} \quad (2.6)$$

$$\sigma^2_{\_j} = \frac{1}{M_{\_j} - 1}\sum_{k=1}^{M_{\_j}}\left(X_{\_jk} - \bar{X}_{\_j}\right)^2. \quad (2.7)$$

The transaction sample size for the *ls* survey is

$$m_{ls} = c_{ls}\sum_{i=1}^{n}\sum_{j=1}^{R} M_{ij}. \quad (2.8)$$

Exhibit 3: Estimator and variance for *es* survey

A two-stage survey using a stand-alone establishment frame is assumed. The stand-alone frame lists the establishments together with the number of their respective transactions $M_{\_j}$. At stage one, a sample of r_{es} establishments is selected by pps with replacement, and at stage two, a fixed size sample c_{es} of transactions is independently selected from the $M_{\_j}$ transactions of establishment j (j = 1, 2, …, r) by srs without replacement.

The unbiased estimator of X is

$$X'_{es} = \left(M/r_{es}\right)\sum_{j=1}^{r_{es}} \bar{X}'_{\_j} \text{ , where} \tag{3.1}$$

$$\bar{X}'_{j} = \sum_{k=1}^{c_{es}} X_{\_jk} \Big/ c_{es}$$

is an unbiased estimate of $\bar{X}_{\_j}$. The variance of X'_{es} is

$$Var\left(X'_{es}\right) = \frac{M^2}{r_{es}}\sigma^2_{es1} + \frac{M}{r_{es}\ c_{es}}\sum_{j=1}^{R}\left(M_{\_j} - c_{es}\right)\sigma^2_{\_j} \text{ ,} \tag{3.2}$$

where

$$\sigma^2_{es1} = \left(1/M\right)\sum_{j=1}^{R} M_{\_j}\left(\bar{X}_{\_j} - X/M\right)^2 \tag{3.3}$$

is the between establishment variance and $\sigma^2_{\_j}$ is the within establishment population variance for establishment j (j = 1, 2, …, R) defined in equation (2.7) in Exhibit 2.

The transaction sample size for the *es* survey is

$$m_{es} = c_{es} r_{es} \,. \tag{3.4}$$

If $c_{es} = c_{ls} = c$ (where c_{ls} is defined in Exhibit 1), and $r_{es} = E\left(\sum_{i=1}^{n}\sum_{j=1}^{R} M_{ij}\right) = n\bar{M}$, the *es* and *ls* transaction sample sizes are equal, $m_{es} = m_{ls}$. Also, if no households have multiple transactions with the same establishment, then $\bar{M} = 1$, $r_{es} = n$, and the difference between the variance of the *ls* and the *es* estimators of X can be written as

$$Var(X'_{es}) - Var(X'_{ls}) = \frac{N^2}{n}\left(\sigma^2_{es} - \sigma^2_{ls1}\right) + \frac{N}{nc}\sum \rho_j \sigma^2_j \text{ ,} \tag{3.5}$$

where the first term and the second terms, respectively, on the right side are the differences between the stage one and stage two variance components of the *ls* and the *es* estimators, and

$$\rho_{\_j} = \left(c/M_{\_j}\right)\sum_{i=1}^{N} M_{ij}\left(M_{ij} - 1\right)\right) \geq 0 \tag{3.6}$$

is the difference between the *es* and the *ls* second stage finite population corrections for establishment j.

amount by which the *ls* second stage variance component is favored depends on the extent to which households have multiple transactions with the same establishments. [See Sirken and Shimizu (circa 2005a).]

D. Comparing *ls* and *ps* sampling errors

The *ls* may be also viewed as a network sampling household survey that counts the transactions of establishments at every household with which the establishments have transactions. This section very briefly summarizes research findings that compare the sampling errors of the *ps* estimator of X and the single-stage and two-stage *ls* estimators of X when the *ps* and *ls* are based on the same population sample survey of households selected by srs WR. The *ps* estimator and variance as shown in Exhibit 4.

The expressions of the *ls* and *ps* estimators of X are equivalent except for the function of the x-variable that is counted when household i has transaction k with establishment j: The *ps* estimator counts X_{jk} and the *ls* estimator counts either $\bar{X}_j$ or $\bar{X}'_j$ depending on whether *ls* is a single-stage or two-stage sample survey. When the within establishment component of variance is ignorable, $X_{\_jk} = \bar{X}_{\_j} = \bar{X}'_{\_j}$, and it follows that sampling errors of the *ps*, and of the one and two-stage *ls* are equivalent. When the within establishment variance component is not ignorable, sampling errors are virtually always greater in the *ps* than in the single stage *ls*.

Also, if the within establishment variance component is not ignorable, sampling errors are virtually always greater in the *ps* than the two stage *ls* whenever sufficiently large *ls* transaction samples are selected in the second stage. For example, if none of the *ls* households is linked to multiple transactions, a condition most likely to be satisfied in surveys of rare populations, equivalent *ls* and *ps* transaction sizes assures that sampling errors are less in the *ls* than in the *ps*. [See Sirken and Shimizu (circa 2005b).]

Exhibit 4: Estimator and variance for *ps* survey

When the *ps* estimator is based on the same population sample survey as the *ls* estimator, the unbiased *ps* estimator of *X* is

$$X'_{ps} = \frac{N}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{R} X_{ij}\right) \quad (4.1)$$

and the variance of X_{ps} is

$$Var(X'_{ps}) = \frac{N^2}{n}\sigma^2_{ps} \text{ where} \quad (4.2)$$

$$\sigma^2_{ps} = (1/N)\sum_{i=1}^{N}\left(\sum_{j=1}^{R} X_{ij} - \bar{X}\right)^2 \quad (4.3)$$

is the population variance between households.

The difference between the variances of the *ps* and *ls* estimators of *X* is

$$Var(X'_{ps}) - Var(X''_{ls}) = \frac{N^2}{n}\left[\sigma^2_{ps} - \sigma^2_{ls1}\right] - \frac{N}{nc}\sum_{i=1}^{N}\sum_{j=1}^{R} M_{ij}\left[1 - \frac{cM_{ij}}{M_{\_j}}\right]\sigma^2_{\_j}. \quad (4.4)$$

Let N^* = the number of households in the population which have transactions ($N^* \leq N$) and let $P = N^*/N$ be the proportion of households having transactions. When none of the N^* households has transactions with multiple establishments and each of them that has transactions with establishment *j* (*j* = 1, 2, …, *R*) has the same number of transactions with establishment *j*, then $M_{ij} = \bar{M}_{\_j} = M_{\_j}/N_{\_j}$ and

$$\sigma^2_{ps} - \sigma^2_{ls1} = \left(1/N^*\right)\sum_{i=1}^{N^*}\sum_{j=1}^{R}\left(X_{ij} - \bar{M}_{\_j}\bar{X}_{\_j}\right)^2. \quad (4.5)$$

If in addition to the earlier conditions, no household has multiple transactions, the difference becomes

$$\sigma^2_{ps} - \sigma^2_{ls1} = \sigma^2_w = (1/M)\sum_{j=1}^{R} M_{\_j}\sigma^2_{\_j} \text{ and} \quad (4.6)$$

$$Var(X'_{ps}) - Var(X''_{ls}) = \frac{PN^2}{n}\sigma^2_w\left(1 - \frac{1}{c}\right) + \frac{PN^2}{nM}\sum_{j=1}^{R}\sigma^2_{\_j} \geq 0,\ c > 0. \quad (4.7)$$

E. Summary

The *ls* is a two-phase hybrid of the traditional population and establishment sample surveys. Phase 1 involves a population sample survey in which household respondents report the number of their transactions with each establishment and identify the establishments. Phase 2 involves surveying the establishments identified in Phase 1 and collecting the variables of interest for their transactions with all households, whether or not Phase 1 households.

This paper summarizes findings from research in which the sampling variance of the *ls* was compared with those of the *es* and the *ps* estimators of the quantity *X*, the *x*-variable summed over the *M* transactions of *N* households with *R* establishments. The research sought to explain the survey differences in sampling errors in terms of the kinds of configurations that are formed by the linkages of *M* transactions between *N* households and *R* establishments. The comparisons assume that the measurement processes in the surveys are flawless.

Conclusions about the relative sampling efficiencies of *ls* and *es*: In single stage sampling, the *ls* and *es* variances of equivalent transactions sample size are equivalent if and only if transactions are uniformly distributed over households such that every household is linked to a single transaction. Deviations from the uniform distribution virtually always increase the *ls* variance and make it less efficient than the *es* variance. The outcome is less clear in two stage sampling. The *ls* and *es* second stage variance components are equivalent if the *es* and *ls* transactions are selected with replacement, but the *ls* second stage variance component is equal to or less than the *es* second stage variance component if *es* transactions are selected without replacement.

Conclusions about the relative sampling efficiencies of *ls* and *ps*: The *ls* and *ps* variances of equivalent household sample size are equivalent if and only if the within establishment component of variance is ignorable. If *ls* is a single stage sample survey, *ls* is likely to be substantially more efficient than *ps*. If *ls* is a two stage sample survey, *ls* is likely to be as efficient or more efficient than *ps* for sufficiently large second stage *ls* samples. For example, if none of the *ls* households has multiple transactions, a second stage *ls* sample size no larger than the number of transactions reported in the survey households is sufficient to assure that the *ls* variance is equal to or less than the *ps* variance.

The research done to date demonstrates that there are potential sampling gains in using the linked population/establishment surveys as a design alternative to the conventional population and establishment surveys. Though these are encouraging findings they are just a beginning. The research virtually ignores effects of differences between *ls* and the traditional surveys (*es* and *ps*) in survey costs, measurement errors, and the utility of the survey data. Nevertheless, hopefully, the methodology and findings will be sufficiently interesting to encourage further design research on linked population/establishment surveys that will ultimately lead to improvements in designing sample surveys to estimate the volume of transactions between the civilian non institutional populations and establishments.

REFERENCES

Wunderlich, G.S. (Ed) (1992). *Toward a National Health Care Survey: a Data System for the 21st Century,* National Research Council and Institute of Medicine, Washington, D.C., National Academy Press.

Sirken M and Shimizu I (circa 2005a). "Establishment Surveys With Population Survey-Generated Sampling Frames." In *Encyclopedia of Biostatistics, Second Edition*, West Sussex, England. John Wiley and Sons Ltd.

—— (circa 2005b). "Establishment based population surveys: design effects of linked population/ establishment surveys compared to conventional population surveys." In *A Handbook of Sampling Methods and Analysis,* P.S.R.S. Rao and M.J. Katzoff, eds. Chapman Hall/CRC Press