Design Effects of Sampling Frames in Establishment Surveys of Small Populations

Monroe Sirken[1], **Iris Shimizu[2]**
[1]Office of Director, National Center for Health Statistics
[2]Office of Research and Methods, National Center for Health Statistics

Abstract

Abstract: This paper compares the design effects of sampling frames in three establishment surveys, LS1, LS2, and ES, of the utilization of establishment services by small populations. Surveys LS1 and LS2 use population survey-generated sampling frames that list establishments which have transactions with people in household sample surveys. In LS1, the population survey-generated sampling frame lists all establishments that have transactions with people in the household sample survey. In LS2, the population survey-generated frame lists the subset of establishments that have transactions with the people in the household sample survey that belong to the small population of interest. Survey ES uses a complete establishment frame that lists all establishments.

KEY WORDS: Establishment Sampling Frames, Small Domain Estimation, Network Sampling

1. Introduction*

This paper summarizes research findings comparing the precision of the Conventional Establishment Survey (ES) and two versions of the Linked Population/Establishment Survey (LS1 and LS2) in estimating a total for a small domain (sd) of the transactions between households and specified kinds of establishments. [1, 2] For example, if the establishments were physicians, the transactions would be visits to physicians. In each comparison, we list the conditions for equivalent precision and discuss how deviations from these conditions may favor one or the other survey. In the ES, a sample of establishments is selected from a complete establishment sampling frame that lists all of the specified kinds of establishments that have transactions with households. In the LS1, a sample of establishments is generated by a population sample survey in which household respondents report all their transactions, including sd transactions, and identify their establishments. In the LS2, the establishment sample is generated by a population sample survey in which household respondents report their sd transactions and identify the establishments with which they had sd transactions.

Precision comparisons assume that a single-stage establishment sample survey is conducted to estimate X, the variable of interest, summed over a small domain of sd transactions of R establishments with N households. Non sampling errors are ignored. Section 2 discusses the three sampling frames considered. ES and LS1 are compared in Section 3 and LS1 and LS2 are compared in Section 4. A summary of the discussions is given in Section 5.

2. Sampling Frames for Establishment Surveys

2.1. Background on Population Generated Establishment Sampling Frames

Two of the considered establishment sampling frames are generated in Linked Population-Establishment Surveys (LSs). In LS, the sampling frame consists of only the establishments reported by the sample of households selected in a population survey. In this type survey, the respondent is first asked to report all of the transactions the household has with establishments. Then, for each reported transaction, the respondent is also asked to provide the name and contact information about the establishment. These establishments (or a sample of them) are then surveyed for information about the variable of interest for transactions they have with all households, not just those which were in the household sample.

Data quality from establishment surveys (either linked population-establishment surveys or conventional establishment surveys) can be better than that from a population survey for items that are better reported by establishments than by households. For example, in a survey about medical care, establishments may provide better information on lab work and other tests performed, prescribed medications, and instructions given the patient because such information may not be completely or accurately reported by patients.

* The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

2.2. Specifics of Three Establishment Sampling frames Considered

Two linked population-establishment surveys (denoted by LS1 and LS2) and a conventional establishment survey (denoted by ES) are considered, here. These surveys use different sampling frames and size measures.

The LS1 frame is an unduplicated listing of establishments reported by households in the population sample survey and the combined number of their transactions with all sampled households. The LS2 frame is an unduplicated listing of establishments reported by sampled households with sd transactions and the combined number of their sd transactions with all sampled households. For example, the sampling frame may be limited to the doctors visited by sampled HHs for the purposes of asthma treatment and the size measures are the numbers of visits those HHs made to each doctor for that treatment.

The ES frame lists all establishments of interest in the universe. Various sources may be used to construct the sampling frames for conventional establishment surveys. For example, phone books and directories may be used for compiling frames of establishments which cater to the public or other establishments. For surveys of health care providers, provider lists may also be obtained from government agencies, such as state health departments, and from professional organizations, such as the American Medical Association and the American Hospital Association. Complete frames may, or may not, list size measures and the measures may, or may not, be correlated with the variable of interest.

3. Compare LS1 and ES estimates for X

3.1. ES Estimator for X

For notation, let:

R and r denote the number of establishments in the universe and the sample, respectively,

$M_j(ES)$ denote the measure of size given in the sampling frame for establishment j (j = 1, 2, … R),

$M(ES) = \sum_{j=1}^{R} M_j(ES)$ = the total frame size,

X_j = the sum of the x-values over all transactions which establishment j (j = 1, 2, …, R) has with HHs,

$X = \sum_{j=1}^{R} X_j$.

If the sample establishments are selected with probability proportional to size (pps) and with replacement, then the Hansen/Hurwitz pps unbiased estimator of X is

$$X'_{ES} = \frac{1}{r_{ES}} \sum_{j=1}^{r_{ES}} \left(X_j / \rho_j \right)$$

where $\rho_j = M_j(ES)/M(ES)$ is the probability of selecting establishment j (j = 1, 2, …, r).

3.2 LS1 Estimator for X

For notation not defined earlier, let:

N and n denote the number of HHs in the universe and the sample, respectively,

M_{ij} = actual number of transactions of HH i with establishment j (i = 1, 2, …, N; j = 1, 2, …, R) (Note that M_{ij} = 0 if HH i has no transactions with establishment j),

$M_{\bullet j} = \sum_{i=1}^{N} M_{ij}$ = actual number of transactions which establishment j (j = 1, 2, …R) has with HHs. This number is the size measure for establishment j in the LS1 sampling frame,

$M = \sum_{j}^{R} M_{\bullet j}$ = total number of transactions which the N households have with the R establishments,

$\bar{X}_j = X_j / \left[M_{\bullet j} \right]$ = is the average of the x-values over all $M_{\bullet j}$ transactions which establishment j has with HHs.

If a simple random sample (srs) of HHs is selected with replacement for the population survey, an unbiased estimate of X may then be expressed as:

$$X'_{LS1} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{R} M_{ij} \bar{X}_j$$

where the M_{ij} is based on information reported by household i (i = 1, 2, …, n) and the $\bar{X}_j$ is based on information reported by establishment j (j = 1, 2, …, r).

To compare the estimators of X from the LS with estimators of X from the ES of equivalent expected establishment sample size, there is a need to first determine the establishment sample size needed from the LS survey. Recall that an establishment is selected to the LS sample each time a transaction with that establishment is reported in the population survey and the number of times (r_{LS1}) establishments are reported in the LS1 sample size is equal to the total number of transactions reported in the LS. That is,

$$r_{LS1} = \sum_{i=1}^{n} \sum_{j=1}^{R} M_{ij}$$

is a random variable with the expected value

$$E\left(r_{LS1}\right) = \left(\frac{n}{N}\right)\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{R} M_{ij} = n\frac{M}{N} .$$

Let $r_{ES} = E(r_{LS1})$.

3.3. Comparison of ES and LS1 estimates for X

When the expected ES sample size equals the expected number of transactions in the LS, the estimates X'_{ES} and X'_{LS1} are unbiased estimates of X and their variances are equivalent if, and only if, three conditions are met:

1. The HHs in the population survey for LS1 are selected with replacement by srs.
2. Each HH has exactly one transaction with establishments. This condition assures there is no clustering of transactions with establishments within individual HHs. The condition also causes the total number of transactions to equal the total number of HHs in the population (i.e. M = N).
3. The ES sample of establishments is selected by pps based on actual establishment size.

Under conditions 1 and 2 the LS1 establishment sample is in essence like an ES sample selected with pps. Deviations from condition 1 due to LS1 complex sample designs and condition 2 due to within household clustering of transactions would favor the ES. Deviations from condition 3 due to poor or absent measures of actual size would favor LS1. If all 3 conditions are not met, it is not clear which of the two surveys is favored.

4. Compare LS1 and LS2 estimates for X

Thus far, we have considered only the linked survey (LS1) in which the establishment frame contains all the establishments that had transactions with households in a population sample survey. This section considers the linked survey LS2 in which the establishment frame is restricted to establishments which have sd transactions with households in the population survey and compares the precision of the LS2 and the LS1 of equivalent sample size.

Asterisks are used where needed in the following to distinguish the notation for LS2 from notation which was defined above for LS1. That is, let:

R* and r* denote number of establishment with one or more sd transactions with HHs in the establishment universe and sample, respectively. These are subsets of the R and r establishments.

M_{ij}^{*} = number of sd transactions HH i has with establishment j (i = 1, 2, …, N; j = 1, 2, …, R).

$M_{\bullet j}^{*} = \sum_{i=1}^{N} M_{ij}^{*}$ = number of sd transactions which establishment j (j = 1, 2, …, R) has with HHs.

$\bar{X}_{j}^{*} = X_{j}^{*} / M_{\bullet j}^{*}$ = average of X values over $M_{\bullet j}^{*}$ sd transactions which establishment j (j = 1, 2, …, R) has with HHs.

If a srs of HHs is selected with replacement for the population survey, an unbiased estimate of X from LS2 may then be formulated as:

$$X'_{LS2} = \frac{N}{n}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{R^{*}} M_{ij}^{*}\bar{X}_{j}^{*} .$$

For illustration, when the sd transactions are limited to those made for the treatment of asthma, the R* and r* establishments are the numbers of doctors in the universe and sample, respectively, who treat patients for asthma, the M_{ij}^{*} is the number of visits made by HH i to doctor j for purposes of asthma treatment, and $\bar{X}_{j}^{*}$ is the average of the X values over all patient visits made to doctor j for asthma treatment.

The estimates X'_{LS1} and X'_{LS2} are equal and their variances are equal if, and only if, every R* establishment has only sd transactions. For example, when the sd transaction is the asthma treatment, this condition means that doctors who treat asthma do not treat any other diagnoses and patients seeking treatment for any other diagnoses than asthma would have to go a doctor other than one of the R* doctors. Under this condition, LS1 and LS2 yield the same number of sd transactions and the same precision. Deviations from this condition would increase the LS1 sd transaction yield and precision relative to LS2 unless the increased yield was clustered in the same households that have sd transactions.

5. Summary

We compared three survey estimators and their precisions for a variable X, the sum of the x variable for transactions of a “small” domain (sd) between households and establishments. The comparisons involve two versions of the Linked Population Establishment Survey (LS) and the Conventional Establishment Survey (ES). The LS uses a household sample survey generated establishment frame which contains the establishments that had transactions with households sampled in a population sample survey. The LS1 frame lists all establishments that have transactions with sample households. The LS2 lists only the establishments that have small domain transactions with households.

Whenever a complete establishment sampling frame has poor coverage or inadequate size measures, the LS deserves serious consideration as a potential design alternative to the ES. This conclusion applies to the LS1 and especially to the LS2 because sd size measures often are unavailable and are not proportional to the total size measures shown in complete establishment sampling frames.

References

[1] Sirken MG (2002). Design Effects of Sampling Frames in Establishment Surveys. *Survey Methodology.* 183-190.

[2] Sirken M and Shimizu I (2005). Establishment Surveys With Population Survey-Generated Sampling Frames. In Armitage P and Colton T, eds. *Encyclopedia of Biostatistics, Second Edition*, Vol. 3. Chichester: John Wiley and Sons, Ltd., 1750-1755.