MORE ON POPULATION BASED ESTABLISHMENT SURVEYS

Iris Shimizu and Monroe Sirken, National Center for Health Statistics (NCHS)
Iris Shimizu, NCHS, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782

Key Words: survey design, network sampling, multiplicity weighting, establishment transactions

1. Introduction[1]

Population based establishment surveys (PBESs) produce statistics about the transactions which household populations have with establishments. For example, a PBES could produce statistics about the numbers and kinds of visits which the household population makes to health care providers. A PBES is a business survey in which sampled establishments are selected through the transactions which households have with them. At the final sampling stage, transactions are selected from those occurring at the sampled establishments. A unique feature of PBESs is that they do not require comprehensive frames from which to select establishment samples. Nor does a self-weighting PBES require independent information about establishment sizes.

A PBES involves three sampling stages. At the first stage a sample of households is selected and sampled individuals are asked to report all of the transactions they made during a specific time period with establishments and then to identify those establishments. At the second stage, sample establishments are selected from a list compiled from only those reported in the household survey. At the third stage, a sample of transactions occurring in each sampled establishment is selected for each transaction reported in the household survey and information is abstracted from the establishment's records for the sampled transactions.

A PBES is a network household survey in which multiplicity counting rules link households that have transactions with the same establishments. Those rules imply that all household transactions with an establishment are countable at every household having any transaction with that establishment. For example, suppose household H_i has transactions with establishments E_1 and E_2. Let M_1 and M_2 be the number of all transactions which establishments E_1 and E_2 have with households. Then the M_1 and M_2 transactions are countable at household H_i.

The PBES differs from the typical network survey in that data about the transactions which are countable in a household are collected from the establishments rather than from the households where they are counted.

This paper discusses several properties of PBES designs with examples taken primarily from the health care industry. Section 2 gives notation needed in later sections. Section 3 presents unbiased PBES estimators while Section 4 presents variances for those estimates when the household samples are selected via simple random sampling. Section 5 compares the variances of the two estimators. Section 6 discusses selected operational aspects associated with implementing PBES designs followed by a summary in section 7.

2. Notation

A population of N households H_i $(i = 1, ..., N)$ has M transactions with L establishments E_j $(j = 1, ..., L)$ during a specified reference period. Let

M_{ij} = number of transactions of H_i with E_j,

then

$$M_{i\cdot} = \sum_j M_{ij} = \text{number of transactions with } H_i\,,$$

$$M_{\cdot j} = \sum_i M_{ij} = \text{number of transactions of } E_j\,, \text{ and}$$

$$M = \sum_i \sum_j M_{ij} = \text{total number of all transactions.}$$

Let X_{jk} represent the variate of interest for the k^{th} $(k = 1, ..., M_{\cdot j})$ transaction with $E_j (j = 1, ..., L)$. The sum of the variate over all M transactions is

$$X = \sum_j \bar{X}_j M_{\cdot j} \tag{1}$$

where

$$\bar{X}_j = \frac{1}{M_{\cdot j}} \sum_k X_{jk}\,. \tag{2}$$

Also let

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

$$\bar{X} = \frac{1}{M} \sum_{j} \sum_{k} X_{jk} \tag{3}$$

= the average value of all transactions with all establishments.

3. PBES Estimator

A PBES network household sample survey is conducted to produce estimate X. It is assumed that the establishment survey is delayed until the household survey is completed and all establishment nominations are matched so that each sampled establishment is visited only once in the survey. The household survey is based on a complex sample design in which n households H_i' (i = 1, ..., n) are selected with probabilities π_i. The survey is based on a counting rule and a subsampling procedure such that each of the M_{ij} transactions of H_i' with E_j (j = 1, ..., L) is linked to a fixed size subsample of transactions independently drawn from the $\sum_{i}^{N} M_{ij} = M_{\cdot j}$ transactions that E_j has with all H_i (i = 1, ..., N).

Let

r index sample transactions selected within E_j ,

c = size of the subsample of E_j's randomly selected transactions that is linked to every transaction of H_i' with E_j , and

X_{jkr} (i) =
information reported about the r^{th} (r = 1, ..., c) randomly selected transaction of E_j in the sample that is linked to the k^{th} (k = 1, ..., M_{ij}) transaction of E_j with H_i'.

Also, let

$$A_i = \{ j | M_{ij} > 0 \}$$

= the subset of the E_j (j = 1, ..., L) for which $M_{ij} > 0$.

In an earlier paper [Sirken, Shimizu, and Judkins (1995)], it was shown that an unbiased estimate of X can be written as:

$$X' = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j \in A_i}^{L} M_{ij} \bar{X}_j'(i) \tag{4}$$

where

$$\bar{X}_j'(i) = \frac{1}{c M_{ij}} \sum_{k}^{M_{ij}} \sum_{r}^{c} X_{jkr}(i) = \frac{1}{t_{ij}} \sum_{r}^{t_{ij}} X_{jr}(i)$$

is an unbiased estimate of $\bar{X}_j$ based on the $t_{ij} = c\ M_{ij}$ transactions randomly selected from E_j because of the transactions reported by H_i'.

Judkins *et al* (in press) also proposed an unbiased estimator for PBES. After some algebra and assuming a single survey period for the household survey, the Judkins, *et al*, estimator can be formulated as:

$$X'' = \sum_{i}^{n} \frac{1}{\pi_i} \sum_{j \in A_i} M_{ij} \bar{X}_j' \ . \tag{5}$$

where

$$\bar{X}_j' = \frac{1}{m_j} \sum_{i}^{n} t_{ij} \bar{X}_j'(i) = \frac{1}{m_j} \sum_{r}^{m_j} X_{jr}$$

is the unbiased estimate of $\bar{X}_j$ based on m_j transactions randomly selected in E_j and

$$m_j = \sum_{i=1}^{n} t_{ij}$$

is the total number of transactions selected from E_j (i.e. across all households which nominated E_j).

The estimators in (4) and (5) are identical except for the sample means used within each establishment. The estimator X' in (4) uses multiple sample means, one for the transaction sample selected for each household nominating that establishment. Hence, for X', one must keep track of the separate transaction samples selected from the establishment for each of the nominating households throughout data collection, processing and estimation procedures. The estimator X'' in (5) requires only the mean of the total sample within each establishment, that is, the sample which results when the transaction samples selected for every household nominating the establishment are combined. That makes X'' simpler operationally than X'.

4. Variances

It is sufficient to consider the variances of PBES estimators X' and X'' for a simple random sample of households selected without replacement. When Ω is a simple random sample of n households in a PBES, then the H_i selection probability is

$$\pi_i = \pi = \frac{n}{N} .$$

The PBES estimate X' in equation (4) becomes

$$X' = \frac{N}{n}\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}\bar{X}_j'(i) \tag{6}$$

Sirken, Shimizu, and Judkins (1995) showed that the variance of (6) can be written as:

$$\sigma^2_{X'} = \frac{N^2}{n}\frac{N-n}{N}\frac{\sum_{i=1}^{N}[X(i)-\bar{X}]^2}{N-1}$$

$$+ \frac{N}{nc}\sum_{i=1}^{N}\sum_{j \in A_i}^{L} M_{ij}\frac{M_{\cdot j}-t_{ij}}{M_{\cdot j}}\sigma^2_j \tag{7}$$

where

$$X(i) = \sum_{j \in A_i} M_{ij}\bar{X}_j \quad \text{and} \tag{8}$$

$$\sigma^2_j = \frac{\sum_{k=1}^{M_{\cdot j}}(X_{jk} - \bar{X}_j)^2}{M_{\cdot j} - 1} \tag{9}$$

is the within establishment variance. The first term of (7) represents the variance contribution due to sampling households and the second term represents the variance contribution due to subsampling transactions that are countable at sample households.

For a simple random sample of households, the PBES estimate X'' in (5) becomes

$$X'' = \frac{N}{n}\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}\bar{X}_j' . \tag{10}$$

An expression for the variance of X'' in (10) can be derived by noting that the variance for an estimator may be written as:

$$\sigma^2_{\hat{X}} = \sigma^2_{E(\hat{X}|\Omega)} + E(\sigma^2_{\hat{X}|\Omega}) , \tag{11}$$

where $(\hat{X}|\Omega)$ denotes the value of the estimate $\hat{X}$ derived from a fixed sample Ω of households. When the sample Ω of households is fixed, the households can be treated as strata and the expected value of X'' in (10) becomes:

$$E(X''|\Omega) = \frac{N}{n}E\left[\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}\bar{X}_j'\right]$$

$$= \frac{N}{n}\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}\bar{X}_j$$

$$= \frac{N}{n}\sum_{i=1}^{n} X(i) \tag{12}$$

where $X(i)$ is defined in (8). The first term of (11) then becomes the well known formula:

$$\sigma^2_{E(X''|\Omega)} = Var\left(\frac{N}{n}\sum_{i=1}^{n} X(i)\right)$$

$$= \frac{N^2}{n}\frac{N-n}{N}\frac{\sum_{i=1}^{N}[X(i) - \bar{X}]^2}{N-1} , \tag{13}$$

which represents the contribution to the variance of X'' due to sampling of households.

Consider the second term of (11) for X''. For a fixed sample of households, the variance of X'' in (10) becomes:

$$\sigma^2_{X''|\Omega} = Var\left[\frac{N}{n}\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}\bar{X}_j'\right]$$

$$= \left(\frac{N}{n}\right)^2\sum_{i=1}^{n}\sum_{j \in A_i}^{L} M_{ij}^2\sigma^2_{\bar{X}_j'} , \tag{14}$$

where

$$\sigma^2_{\bar{X}_j'} = Var\left(\frac{1}{m_j}\sum_{k=1}^{m_j} X_{jk}\right) = \frac{M_{\cdot j} - m_j}{m_j M_{\cdot j}}\sigma^2_j$$

$$= \left(\frac{1}{m_j} - \frac{1}{M_{\cdot j}}\right)\sigma^2_j \tag{15}$$

and σ^2_j is the within establishment variance in (9). If we let both i and i' index the sample households, then for each sample H_i', the m_j can be reformulated as

$$m_j = \sum_i^n t_{ij} = c\left(M_{ij} + \sum_{i' \neq i}^n M_{i'j} \right)$$

so that (15) can be written as:

$$\sigma^2_{\bar{X}_j'} = \left(\frac{1}{c} \frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} - \frac{1}{M_{\cdot j}} \right) \sigma_j^2 \ . \qquad (16)$$

Using (14) and (16), the second term of (11) then becomes

$$E(\sigma^2_{X''|\Omega}) \doteq E\left[\left(\frac{N}{n} \right)^2 \sum_{i=1}^n \sum_{j \in A_i} M_{ij}^2 \sigma^2_{\bar{X}_j} \right]$$

$$= \frac{N}{n} \sum_{i=1}^N \sum_{j \in A_i} M_{ij}^2 \left(\frac{1}{c} E\left[\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'i}} \right] - \frac{1}{M_{\cdot j}} \right) \sigma_j^2 \qquad (17)$$

which represents the contribution to the variance of X'' due to the sampling of transactions within establishments.

Using (13) and (17) in (11), the vàriance of X'' is thus:

$$\sigma^2_{X'} = \frac{N^2}{n} \frac{N-n}{N} \frac{\sum_{i=1}^N [X(i) - \bar{X}]^2}{N-1}$$

$$+ \frac{N}{n} \sum_{i=1}^N \sum_{j \in A_i} M_{ij}^2 \left(\frac{1}{c} E\left[\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} \right] - \frac{1}{M_{\cdot j}} \right) \sigma_j^2 \ . \qquad (18)$$

5. Comparing variances

It can be seen that the first terms of variances (7) and (18) for X' and X'', respectively, are identical. Hence the difference between the variances for PBES estimators X' and X'' is the difference between second terms of variances for the two estimators. That is, after some algebra, the difference in variances becomes:

$$\sigma^2_{X''} - \sigma^2_{X'}$$

$$= \frac{N}{n} \sum_{i=1}^N \sum_{j \in A_i} M_{ij}^2 \left(\frac{1}{c} E\left[\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} \right] - \frac{1}{M_{\cdot j}} \right) \sigma_j^2$$

$$- \frac{N}{nc} \sum_{i=1}^N \sum_{j \in A_i} M_{ij} \frac{M_{\cdot j} - cM_{ij}}{M_{\cdot j}} \sigma_j^2$$

$$= \frac{N}{nc} \sum_{i=1}^N \sum_{j \in A_i} M_{ij}^2 \left(E\left[\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} \right] - \frac{1}{M_{ij}} \right) \sigma_j^2 \ . \qquad (19)$$

In (19), the summation over establishments for a specific household i is limited to establishments with which the household has transactions so that $M_{ij} > 0$. Thus

$$\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} \leq \frac{1}{M_{ij}} \ .$$

and

$$E\left[\frac{1}{M_{ij} + \sum_{i' \neq i}^n M_{i'j}} \right] \leq \frac{1}{M_{ij}} \qquad (20)$$

for every E_j with which H_i has transactions. Using (20) in (19), we get:

$$\sigma^2_{X''} \leq \sigma^2_{X'} \ . \qquad (21)$$

Hence the PBES estimator X'' which uses a single sample mean from each sample establishment is not only easier operationally, it is also more precise than the estimator which uses multiple sample means from each sample establishment. This result has implications for PBESs which use cluster samples of households where clusters are defined geographically because clustered households are more likely to have transactions with the same establishments and, hence, are more likely to yield multiple nominations for each establishment than are households in a simple random sample.

6. Operational Aspects

We now look at operational aspects one should also consider when designing a PBES.

6.1 Required Volume of Industry Transactions

In PBES the household survey must generate the number of sample establishments specified for meeting the objectives of the PBES. The targeted establishment universe must have a fairly large number of transactions with households. For example if the National Health

Interview Survey (NHIS) sampling weights were approximately 1,900 and if a 4-week reference period were used, it has been speculated that about 10 million provider visits annually would be required in order for the NHIS household sample to generate a national sample of about 400 health care providers. (Judkins, *et al* in press) Alternatively, in today's world, industries with so many transactions may have membership lists which could be used as sampling frames in a list sample survey.

6.2. Costs

A PBES may be more expensive than a list sample of establishments because it requires use of a household survey to construct a sampling frame of establishments that will be surveyed. That is, resources must be added in a household survey to collect and process information about the transactions which households have with establishments and about the establishments for those transactions. Such expenditures are not required if a list were available for sampling. The expenses are even greater if one does not have a household survey being conducted for other purposes.

On the other hand costs for a PBES could be less than those of a list sample if one is already conducting a followback study in which establishments having transactions with households are identified and visited. A PBES could be conducted in those establishments which are part of the other survey, thus eliminating the costs of compiling an establishment sampling frame. If the other study requires travel to those establishments, travel costs for the PBES could also be reduced, if not eliminated, by conducting the PBES in the establishments at the same time as the other study.

6.3. Timing

From start to finish, a PBES survey is likely to require a longer time to complete than a list sample survey of establishments because in a PBES one must allow time for a household survey to generate nominations of sample establishments before fielding the establishment part of the survey. If there are no existing household surveys which can be used by the PBES, additional time will be needed to initiate the household survey. One can minimize the time to complete a PBES by starting field work on the establishment survey as soon as some establishment nominations are received from the household survey. However, if one is not pressed for time, the results derived in Section 5 indicate that one may improve the precision in PBES estimates by delaying the establishment survey until the household survey is complete and duplicate nominations for individual establishments are matched so that all transaction samples can be selected at the same time in each establishment that is nominated multiple times. Conducting all data collection at a single time in each establishment also offers the benefit of minimizing response burden and, hence, the risk of refusals in establishments that are hit multiple times.

6.4. Household Response Errors

One must design surveys to minimize the effect of response errors which can affect the calculation of the weights for sample establishments. In particular, there is concern about response errors known to occur in the household surveys.

Respondents may fail to report transactions for a variety of reasons. A respondent may refuse to report transactions that are sensitive in nature. For example, one may not want to report visits made to health care providers for treatment of HIV. If respondents are asked to report for others in their households, the respondent may not know about all of the transactions which other household members had with establishments. For example a wife may not know her husband went to a doctor. Respondents may forget some transactions entirely, especially, if there are many of them. There are also telescoping problems in which respondents erroneously place transactions inside, or outside, of the referenced time period. These errors happen in both traditional and network household surveys.

Some response errors resulting in over- or under reporting of transactions may be identified by asking the nominated establishments about the transactions which they had with the sample households during the survey reference period. Transactions reported by the households could then be compared with those listed by each establishment to identify the falsely reported transactions and to identify transactions that did occur but which were not reported by the household. Usefulness of such a follow-back survey would probably be reduced for some establishment types, such as health care providers, which would restrict their transaction disclosures to those for households which give the establishment consent to divulge that information. Regardless of establishment type, however, follow-back studies could never detect unreported transactions that occurred with establishments which were never nominated by any household respondent.

The counting rules must be designed to minimize affects due to respondent misclassification of establishments with which they have transactions. For example, suppose one wants to use PBES to estimate the number of transactions with internists but many respondents think their internists are family practitioners. The respondents would thus erroneously omit reporting transactions with the internists if asked only for transactions with internists. To resolve this shortfall, one may broaden the reporting rule to include family and general practitioners as well as internists, then respondents would report transactions to all providers whom they think fall in those categories. Unless one also wants to estimate transactions for the other provider types, one must plan to screen out the unwanted providers during the establishment survey.

One must develop methods to maximize capturing information required in PBES to locate the establishment for each reported transaction. Survey experience, such as in the National Medical Expenditure Survey [Johnson (1995)], shows that substantial portions of respondents may not provide enough identifying information to enable contact with reported establishments. The establishments not contacted are treated as non-respondents in a PBES.

When an establishment's transactions with a sample household are not reported or, if when reported, the transactions are not linked correctly to that establishment, the survey weights for that establishment are erroneously reduced and the resulting PBS estimates are reduced. Over-reporting of transactions contributes to over-statement in the PBES estimates or to added survey costs if steps are made to eliminate the erroneously reported transactions.

7. Summary

A population based establishment survey (PBES) is a method to consider when one wants to conduct a survey of establishments for which no list exists nor can be constructed. A PBES is a household sample survey which can produce unbiased estimates. However, experimentation is needed to improve the design of such surveys.

References

Johnson, A. (1995). "Business Surveys as a Network Sample." *Business Survey Methods*, Edited by Cox, Binder, Chinnappa, Christianson, Colledge, Kott. John Wiley & Sons, Inc. pp 219-233.

Judkins, D.; Berk, M.; Edwards, S; Mohr, P.; Stewart, K.; and Waksberg, J. "National Health Care Survey: List Versus Network Sampling." National Center for Health Statistics. *Vital Health Stat 2* (in press).

Sirken, M.; Shimizu, I.; Judkins, D. (1995). "The Population Based Establishment Surveys." *Proceedings of the Survey Research Section, American Statistical Association*. pp 470-473.