# THE POPULATION BASED ESTABLISHMENT SURVEYS

**Monroe Sirken, Iris Shimizu, National Center for Health Statistics (NCHS); David Judkins, Westat, Inc.**
**Monroe Sirken, NCHS, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782**

**Key Words: network sampling, multiplicity weighting, establishment transactions**

## I. Introduction

The Population Based Establishment Survey (PBES) is an establishment sample survey that compiles statistics on transactions between populations and establishments. It is a unique establishment survey because it is based on a sample of establishments that had transactions with, and were reported by households interviewed in population sample surveys. Contrary to the sample design of conventional establishment surveys, PBES does not require a comprehensive free-standing establishment frame with good measures of establishment size. The conditions favoring PBES over the conventional establishment sample survey or vice versa will depend largely on the comparative costs and quality of their respective sampling frames. Since very little is known about the statistical properties of PBES designs, the National Center for Health Statistics has been collaborating with Westat, Inc. in investigating some of the nonsampling and sampling issues (Judkins, et al., 1995). However, this paper's objective is more modest. It describes the PBES sample design and presents the unbiased PBES estimator and its sampling variance.

## II. PBES Sample Design

PBES is a network household survey (Birnbaum and Sirken, 1965; Sirken, 1970; Sirken, 1972) that is based on a 2-phase sample design in which households are the first phase sampling units and the second phase sampling units are the transactions of establishments that can be nominated by sample households.

The transactions eligible for nomination by sample households are determined by the PBES multiplicity counting rule (Sirken, 1974) which states that "households are eligible to nominate all transactions of establishments with whom they have had transactions". The transactions that can be nominated are subsampled with sample sizes proportional to the number of transactions that households had with the establishments.

For example, consider sample household, $H_i$. It had two transactions with establishment $E_1$ and three transactions with establishment $E_2$. Establishments $E_1$ and $E_2$ had a total of $M_1$ and $M_2$ transactions respectively with all households in the population, whether or not included in the survey sample. Then, in compliance with the PBES multiplicity rule, $H_i$ is eligible to nominate a random subsample of $2c$ of $E_1$'s transactions and $3c$ of $E_2$'s transactions, where $c$ is a positive integer.

Though subsamples of establishments' transactions can be nominated by sample households having transactions with them, information about these transactions is reported by establishments and not by households. Sample households identify the establishments with whom they had transactions and they report the number of their transactions with each of them. The information about these transactions is subsequently collected in surveys of the establishments previously identified by sample households.

There are essentially three PBES implementation phases. Phase 1 involves a network household sample survey in which sample households identify establishments with whom they had transactions and report the number of their transactions with each of them. Phase 2 involves compiling a frame of all establishments that had transactions with and were reported by households in the sample survey. For each listed establishment, the frame indicates the total number of its transactions with sample households. Phase 3 involves a survey with the frame's establishments. Transactions are subsampled with sample sizes proportional to the number of transactions the establishment had with all sample households.

## III. 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 calendar period. Let

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

then

$M_{i.} = \sum_j M_{ij}$ = number of transaction with $H_i$

$M_{.j} = \sum_i M_{ij}$ = number of transactions of $E_j$

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

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

$$X = \sum_j \bar{X}_j M_{.j}$$

where

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

Also let

$$\bar{X} = \frac{1}{M} \sum_j \sum_k X_{jk}$$

= the average value of all transactions with all establishments.

## IV. PBES Estimator

A PBES network household sample survey is conducted to estimate $X$. The household survey is based on a complex sample design in which $n$ households $H_i'$ ($i = 1, ..., n$) are selected with probabilities $\pi_i$. The survey is based on a counting rule 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 $\sum_i^N M_{ij} = M_{.j}$ transactions that $E_j$ has with all $H_i$ ($i = 1, ..., N$).

Let

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 $r^{th}$ ($r = 1, ..., l$) 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 $M_{ij} > 0$.

A linear estimate of $X$ in terms of the $X_{jkr}(i)$ is

$$X' = \sum_i^n \sum_{j \in A_i}^r \sum_r^c \frac{W_{jk}}{\pi_i} X_{jkr}(i) \qquad (1)$$

which is unbiased if, and only if,

$$\sum_i^N \sum_{j \in A_i}^L \sum_k^{M_{ij}} W_{jk} c \bar{X}_j = \sum_i^N \sum_j^L \sum_k^{M_{ij}} \bar{X}_j . \qquad (2)$$

Now (2) is an identity in $\bar{X}_j$ if, and only if,

$$W_{jk} = \frac{1}{c} \qquad (3)$$

where the $W_{jk}$'s are multiplicity weights assigned to the transactions of $E_j$. Substituting (3) in (1),

$$X' = \sum_i^n \frac{1}{\pi_i} \sum_{j \in A_i} \sum_k^{M_{ij}} \frac{1}{c} \sum_r^c X_{jkr}(i)$$

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

where

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

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

Finally, the unbiased estimate of $X$ based on the PBES sample of n households can be written as

$$X' = \sum_{i=1}^{n} \frac{X'(i)}{\pi_i} \tag{5}$$

where

$$X'(i) = \sum_{j \in A_i} M_{ij}\ \bar{X}_j'(i)$$

is an unbiased estimate of $X(i) = \sum_{j} M_{ij}\ \bar{X}_j$ .

## V. PBES Variance

It is sufficient to derive the variance of $X'$ for a simple random sample of households selected without replacement. The variance of estimate $X'$ may be written as:

$$\sigma^2_{X'} = \sigma^2_{E(X'|\Omega)} + E(\sigma^2_{X'|\Omega}) , \tag{6}$$

where $(X'|\Omega)$ denotes the value of the estimate $X'$ derived from a fixed sample $\Omega$ of households. When $\Omega$ is a simple random sample of n households in a PBES, then

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

and the PBES estimate in equation (4) becomes

$$X' = \frac{N}{n} \sum_{i=1}^{n} \sum_{j \in A_i}^{L} M_{ij} \bar{X}_j'(i) . \tag{7}$$

For a fixed sample $\Omega$ of households, the households can be treated as strata and the expected value of $X'$ becomes:

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

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

$$= \frac{N}{n} \sum_{i=1}^{n} X(i) . \tag{8}$$

If we let

$$\bar{X} = \frac{\sum_{i=1}^{N} X(i)}{N} , \tag{9}$$

the first term of (6) then becomes the well known formula:

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

$$= \frac{N^2}{n} \frac{N - n}{N} \frac{\sum_{i=1}^{N} [X(i) - \bar{X}]^2}{N - 1} , \tag{10}$$

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

Consider the second term of (6). For a fixed sample of households, the variance of $X'$ in (7) becomes:

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

$$= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{n} \sum_{j \in A_i}^{L} M_{ij}^2 \sigma^2_{\bar{X}_j'(i)} , \tag{11}$$

where

$$\sigma^2_{\bar{X}_j'(i)} = Var\left(\frac{1}{t_{ij}} \sum_{k=1}^{t_{ij}} X_{jk}\right) = \frac{M_{\cdot j} - t_{ij}}{t_{ij} M_{\cdot j}} \sigma_j^2 \tag{12}$$

and

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

is the within establishment variance. Because $t_{ij} = c\, M_{ij}$, (12) can also be written as:

$$\sigma^2_{\bar{X}'_{(i)}} = \frac{M_{\cdot j} - t_{ij}}{c M_{ij} M_{\cdot j}} \sigma_j^2 \quad . \tag{13}$$

The second term of (6) then becomes

$$E(\sigma^2_{X'|\Omega}) = \left(\frac{N}{n}\right)^2 E\left(\sum_{i=1}^{n} \sum_{j \in A_i}^{L} M_{ij}^2 \frac{M_{\cdot j} - t_{ij}}{c M_{ij} M_{\cdot j}} \sigma_j^2\right)$$

$$= \frac{N}{nc} \sum_{i=1}^{N} \sum_{j \in A_i}^{L} M_{ij} \frac{M_{\cdot j} - t_{ij}}{M_{\cdot j}} \sigma_j^2 \,, \tag{14}$$

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

Using (10) and (14) in (6), the variance of $X'$ is thus:

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

$$+ \frac{N}{nc} \sum_{i=1}^{N} \sum_{j \in A_i}^{L} M_{ij} \frac{M_{\cdot j} - t_{ij}}{M_{\cdot j}} \sigma_j^2 \,. \tag{15}$$

## VI. Concluding Remarks

PBES represents the latest stage in the evolution of a survey design research program that the National Center for Health Statistics initiated about 20 years ago. The research program's objective is to integrate the sample designs of NCHS' independently designed national sample surveys. The integration strategy adopted involves linking NCHS' National Health Interview Survey (NHIS) (Sirken and Greenberg, 1983) to the other NCHS surveys. Thus, the Center's population surveys, such as the National Survey of Family Growth, are being integrated by generating their sample's households and persons from those enumerated in NHIS. The Center's establishment surveys, including four out of six of its medical provider surveys, are being integrated geographically by selecting provider samples within subsets of NHIS PSU's.

More recently, a Panel of the Committee on National Statistics (Wunderlich, 1992) in reviewing NCHS plans for integrating its family of health provider surveys recognized that linking the health provider surveys to NHIS would be more meaningful if the linkage occurred at the medical provider level rather than at the aggregated PSU level. Hence, the Panel proposed that NCHS investigate the feasibility of using listings of providers that had transactions with and were reported by NHIS sample households as sampling frames for the medical provider surveys. The Panel's proposal served as a catalyst in initiating research on PBES sample designs, including the work on deriving the unbiased PBES estimator and its variance that is presented in this paper.

## References


Birnbaum, Z.W.; Sirken, M.G. (1965). "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases." *Vital and Health Statistics*. National Center for Health Statistics, Washington, D.C.

Judkins, D.; Berk, M.; Edwards, S; Mohr, P.; Stewart, K.; and Waksberg, J. (1995). "National Health Care Survey: List Versus Network Sampling." Unpublished report. National Center for Health Statistics.

Sirken, M.G. (1970). "Household Surveys with Multiplicity." *Journal of American Statistical Association*, 65, pp 257-266.

Sirken, M.G. (1972). "Variance Components of Multiplicity Estimators." *Biometrics*, 22, pp 869-873.

Sirken, M.G. (1974). "The Counting Rule Strategy in Sample Surveys." *Proceedings of the Social Sciences Section of the American Statistical Association*, pp 119-123.

Sirken, M.G.; Greenberg, M.S. (1983). "Redesign and Integration of a Population Based Health Survey Program." *Proceedings of the 44th Session of the International Statistical Institute.*

Wunderlich, G.S. (ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. National Academy Press. Washington, D.C.