THE HORVITZ-THOMPSON ESTIMATOR IN POPULATION BASED ESTABLISHMENT SAMPLE SURVEYS

Monroe Sirken and Iris Shimizu, National Center for Health Statistics
Monroe Sirken, National Center for Health Statistics, 6525 Belcrest Road, Room 700, Hyattsville, MD 20782

Key Words: network sampling; establishment transactions; integrated sample design

1. Introduction[1]

Whenever free-standing sampling frames are unavailable or when available frames lack good coverage of establishments or lack good measures of establishment size, the Population Based Establishment Survey (PBES) is an attractive design alternative to the conventional establishment sample survey. And whenever the variate of interest refers to rare and elusive populations that are hard to reach directly, the PBES is an attractive design alternative to the conventional population sample survey.

This paper presents the PBES Horvitz-Thompson estimator of X, the sum of a variate over the M transactions of R establishments. Let M_j be the total number of transactions of the E_jth (j = 1, ..., R) establishment during a specified calendar period. The task at hand is to design a multipurpose establishment sample survey to estimate the Xs for a large number of different variates. Typically, establishment surveys that seek to estimate X are designed as two-stage sample surveys in which establishments are selected with probabilities proportionate to size, and their transactions are the second stage selection units. Designed in this manner, establishment surveys require free-standing sampling frames with good coverage of R establishments and good measures of establishment sizes, the M_js.

Though listings of households and persons enumerated in population sample surveys often serve as sampling frames for other population sample surveys (Mathiowetz, 1987; Cox, Folsom and Virage, 1987), listings of establishments that have transactions with households in population sample surveys rarely serve as frames for establishment sample surveys. The Consumer Price Index (CPI) which depends on data collected in population and establishment surveys (Leaver and Valliant, 1995) is a notable exception. Households enumerated in the CPI Continuing Point of Purchase Survey (CPOPS), a population sample survey, report the establishments with whom they had transactions (purchased merchandise). The listing of establishments reported in CPOPS serves as the sampling frame for the CPI Pricing Survey, a sample survey of retail establishments that collects prices for a basket of consumer goods.

Several years ago, a Panel of the Committee on National Statistics, National Research Council, (Wunderlich, 1992), suggested that the National Center for Health Statistics (NCHS) investigate the feasibility and potential gains of using listings of medical providers that are reported by households in the National Health Interview Survey (NHIS) as sampling frames for NCHS's national medical provider sample surveys which were then and still are independently designed as conventional establishment sample surveys. [The NHIS is an on-going household survey of about 42,000 households annually that is conducted by the NCHS to obtain national health statistics for the U.S. civilian non-institutional population (Massey, Moore, Parsons, and Tadros, 1989)]. The Committee's suggestion initiated a PBES research program at NCHS.

Judkins, Berk, Edwards, Mohr, Stewart and Waksberg (1995) compared the operational and design features of the health care surveys if linked to NHIS with design features of independently designed health care surveys. Judkins, Marker, Waksberg, Botman and Massey (1999) made rough cost/error comparisons of an independently designed dental survey and a dental survey linked to NHIS. They tentatively concluded that if a reasonable list with a reasonable measure of size can be found, an independently designed dental survey is probably preferable, and otherwise the dental survey linked to NHIS should be considered.

More recently, the PBES research has been theoretically oriented, focusing on the problem of constructing alternative unbiased PBES estimators with different data requirements, and getting closed formulas for their variances. Conceptual difficulties initially encountered in this effort were overcome once it was recognized that the PBES is a population network sample survey (Sirken, 1970). Applying network sampling theory, Sirken, Shimizu, and Judkins (1995), and Shimizu and Sirken (1998) obtained two versions of the unbiased PBES multiplicity estimator and derived their variances. In this paper, we present the unbiased PBES Horvitz-Thompson estimator and its variance. The PBES estimators are essentially extensions to multiple stage sampling under special conditions of single-stage

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

network sampling estimators that were originally proposed by Birnbaum and Sirken (1965), and described by Thompson (1992).

2. Notation

Let M_j represent the number of transactions of the E_jth (j = 1, ..., R) establishment. Then

$$M = \sum_{j=1}^{R} M_j = \text{the total number of transactions of the R establishments.} \qquad (1)$$

Let N_j = the number of households having transactions with E_jth (j = 1, ..., R) establishment, N_{jl} = number of households having transactions with both E_jth and E_lth (j $\neq l$) establishments, and N_0 = number of households not having any transactions with any establishments. Then

$$N^* = \sum_{j=1}^{R} N_j - \sum_{j \neq l}\sum N_{jl}$$
$$= \text{the total number of households having transactions with R establishments,} \qquad (2)$$

and

$$N = N^* + N_0 = \text{the total number of households.} \qquad (3)$$

Let the value of the variate for the kth (k = 1, ..., M_j) transaction of the E_jth (j = 1, ..., R) establishment be denoted by X_{jk}. Then

$$X_j = \sum_{k=1}^{M_j} X_{jk} = \text{sum of the variate over } M_j \text{ transactions of the } E_j\text{th establishment,} \qquad (4)$$

and

$$X = \sum_{j=1}^{R} X_j = \text{the sum of the variate over M transactions of R establishments.} \qquad (5)$$

3. The Network Sampling Error Model

A PBES is conducted to estimate X. First, a population sample survey based on a random sample of n households H_i(i = 1, ..., n) is conducted in which sample households identify each of the establishments with whom they had transactions during a specified calendar period. After eliminating duplicate reports of the same establishments, a follow-on establishment survey is conducted with the r distinct establishments reported by n households in the population sample survey, and each sample establishment E_j (j = 1, ..., r) independently selects and reports the variates for a random sample m_j of its M_j transactions.

Judkins et. al. (1999) view the PBES as a 2-stage *establishment* sample survey in which the r establishments that had transactions with n sample households in the population survey are first stage selection units, and the m_j transactions (j = 1, ..., r) selected by each of the r establishments, are second stage sampling units. However, the PBES design features become more transparent, and the PBES estimators and their variances more tractable when the PBES is modeled as a 2-stage network sample *population* survey. From the network sampling perspective, households are first stage units, and transactions that are countable at sample households in compliance with the PBES counting rule are second stage units.

The PBES counting rule specifies that every household in the network of N_j households that had transactions with E_j (j = 1, ..., R) is linked to the same fixed size random sample m_j of the M_j transactions of the E_j establishment. The PBES counting rule implies that the m_j transactions of E_j (j = 1, ..., R) are countable in the population survey at every sample household belonging to the network of N_j households that had transactions with E_j. From the network sampling perspective, establishments that have transactions with households are proxy respondents for transactions that are countable at households. PBES households do not report about their own transactions nor about the transactions countable at their addressees vis-a-vis the PBES counting rule. Households identify establishments with whom they had transactions and those establishments select the subsamples of their transactions that are countable at sample households and they report the variates for the selected transactions.

The PBES counting rule produces a configuration of transactions between establishments and households that partitions the N households into R establishment networks, A_j (j = 1, ..., R), where the A_jth network contains the set of N_j households and is linked to the M_j transactions of E_j. Though the same household may belong to multiple networks, each of the M transactions is uniquely linked to one and only network.

Networks are counted differently by PBES multiplicity estimators and by the Horvitz-Thompson estimator. Multiplicity estimators count the M_j transactions linked to the A_jth (j = 1, ..., R) network every time households belonging to the A_jth network are selected in the population survey sample. The Horvitz-Thompson estimator does not depend on the number of times that households belonging to the same networks are selected in the population survey. The PBES Horvitz-Thompson estimator counts each distinct network only once.

4. The PBES Horvitz-Thompson Estimator

For a sample of n households selected by simple random sampling, and a total sample of

$$m = \sum_{j=1}^{r} m_j \text{ transactions,} \qquad (6)$$

where the transaction subsamples m_j (j = 1, .., r) are selected independently and by simple random sampling, the PBES Horvitz-Thompson estimator of X is

$$X' = \sum_{j=1}^{R} \frac{\alpha_j}{p_j} X_j' \; . \qquad (7)$$

Here, α_j is a random variable that is equal to 1 if any of the n sample households belongs to the A_jth network and α_j is equal to 0 otherwise, and

$$X_j' = M_j \sum_{k=1}^{m_j} \frac{X_{jk}}{m_j} \qquad (8)$$

is the unbiased estimator of X_j (j = 1, .., R) and

$p_j = E(\alpha_j) =$ the probability of any of the n sample households belonging to A_jth (j = 1, ..., R) network. (9)

X' is an unbiased estimate of X if everyone of the R establishments has transactions with at least one household.

Let

$q_j = 1 - p_j =$ the probability that none of the n sample households belongs to the A_jth network. (10)

If n households are selected by simple random sampling without replacement,

$$q_j = \frac{\binom{N - N_j}{n}}{\binom{N}{n}} \; . \qquad (11)$$

If n households are selected by simple random sampling with replacement,

$$q_j = \frac{(N - N_j)^n}{N^n} . \qquad (12)$$

There are two potential measurement problems involving the $q_j s$ (j = 1, ..., r). First, they are dependent on the $N_j s$ (j = 1, ..., r), quantities that are often difficult to ascertain in establishment surveys. Second, it would be difficult to compute the $q_j s$ for most population surveys which, like the NHIS, are based on complex sample designs.

5. The Variance of the PBES Horvitz-Thompson PBES Estimator

The variance of the Horvitz-Thompson estimator of X may be written as

$$Var(X') = VarE(X'|\Omega) + E(VarX'|\Omega) \qquad (13)$$

where $(X'|\Omega)$ denotes the value of X' conditional on a fixed sample Ω of n households.

Consider the first term on the right side of (9),

$$VarE(X'|\Omega) = \text{Var}\left(\sum_{j=1}^{R} \frac{\alpha_j X_j}{p_j} \right)$$

$$= \sum_{j=1}^{R} \frac{X_j^2}{p_j^2} Var(\alpha_j)$$

$$+ \sum_{j=1}^{R} \sum_{l \neq j} \frac{X_j}{p_j} \frac{X_l}{p_l} Cov(\alpha_j \alpha_l). \qquad (14)$$

Since α_j is a binomial random variable

$$\text{Var}\,(\alpha_j) = p_j - p_j^2 \qquad (15)$$

and

$$\text{Cov}\,(\alpha_j \alpha_l) = p_{jl} - p_j p_l \qquad (16)$$

where

$$p_{jl} = 1 - q_j - q_l + q_{jl}^* \quad (j \neq l) \qquad (17)$$

is the joint probability that any of the n sample households belongs to the A_jth and the A_lth networks, and $q_{jl}^*(j \neq l)$ is the probability that the n sample households are linked to neither the A_jth nor A_lth network.

For simple random sampling of n households with replacement,

$$q_{jl}^* = \frac{(N - N_j - N_l + N_{jl})^n}{N^n}, \qquad (18)$$

and for simple random sampling of n households without replacement,

$$q_{jl}^* = \frac{\binom{N - N_j - N_l + N_{jl}}{n}}{\binom{N}{n}} \; . \qquad (19)$$

Consider the second term on the right side of (13),

$$E\,(\mathrm{Var}\,X'\,|\,\Omega) = E\left[\sum_{j=1}^{R} \frac{\alpha_j}{p_j^2} M_j^2 Var\,(\bar{X}_j')\right]$$

$$= \sum_{j=1}^{R} M_j^2 \frac{Var(\bar{X}_j')}{p_j} \ . \qquad (20)$$

$$\mathrm{Var}\,(\bar{X}_j') = \frac{M_j - m_j}{m_j M_j} \sigma^2(X_j) \qquad (21)$$

where the population variance

$$\sigma^2(X_j) = \frac{\sum_{k=1}^{M_j} (X_{jk} - \bar{X}_j)^2}{M_j - 1} \ . \qquad (22)$$

Optimum allocation of m transactions to minimize the variance in (20) is achieved with the establishment subsample sizes

$$m_j = m \frac{\sigma_j M_j / \sqrt{p_j}}{\sum_{j=1}^{R} \sigma_j M_j / \sqrt{p_j}} \ . \qquad (23)$$

Thus, the optimization allocates larger sample sizes to the large and more variable establishments having small selection probabilities.

Combining (14) and (20), the variance of the PBES Horvitz-Thompson estimator of X is

$$\mathrm{Var}\,(X') = \sum_{j=1}^{R} \frac{1 - p_j}{p_j} X_j^2$$

$$+ \sum_{j=1}^{R} \sum_{l \neq j} \frac{p_{jl} - p_j p_l}{p_j p_l} X_j X_l$$

$$+ \sum_{j=1}^{R} \frac{M_j^2}{p_j} \frac{M_j - m_j}{m_j M_j} \sigma^2(X_j). \qquad (24)$$

The first two terms on the right side of (24) represent the between establishment component of variance due to sampling households. The second term on the right vanishes if none of the N households has transactions with more than one establishment. The third term on the right side of (24) is the within establishment component of the variance due to subsampling transactions, and vanishes in single stage sampling when the sample establishments report the variates for all their transactions. Single stage sampling is more likely to be the design option in a single purpose PBES than in a multi- purpose PBES, especially when the variate of interest represents a relatively rare event.

6. Concluding Remarks

All unbiased PBES estimators, whether the Horvitz-Thompson estimator proposed in this paper or the PBES multiplicity estimators proposed by Sirken, Shimizu, and Judkins (1995) and Shimizu and Sirken (1998) depend on multiplicity parameters to adjust for variations in the selection probabilities of the establishments reported in the population sample survey. However, multiplicity and Horvitz-Thompson estimators differ in the ways multiplicities are defined and in likelihood of successfully collecting this information in the follow-on survey with the establishments that were reported in the population survey.

The feasibility and ease with which establishments can provide the multiplicity information is a key factor in deciding on which kind of PBES estimator, if any, is most appropriate in particular applications. The $N_j s$ and $M_j s$ (j = 1, ..., r) respectively are the multiplicities needed by the PBES Horvitz-Thompson estimator and the PBES multiplicity estimators, where N_j is the number of households having transactions with the E_jth establishment, and M_j is the total number of transactions of the E_jth establishment. The $N_j s$ are unlikely to be readily available except at establishments, such as health maintenance organizations, utility companies, and home owner insurance companies, for which households are the transactional units. On the other hand, the $M_j s$ are likely to be available at many establishments that tend to keep track of the total number of services provided though unlikely to know the number of households to whom services were provided.

The PBES is a sample survey design option with many potential applications. It is a mechanism for linking population sample surveys to data files of establishments. Because the mechanism does not require disclosure of personal identifiers, PBES would not be restricted by the kinds of confidentiality concerns that ordinarily limit access to establishment data files. PBES offers the prospects of being able to estimate the volume of establishment transactions under circumstances beyond the capabilities of conventional establishment sample surveys when free-standing establishment frames are unavailable or inadequate, and beyond the capabilities of conventional population sample surveys when the variates of interest relate to rare and elusive populations that are hard to reach directly. Determining which, if any, of these and possibly other potential PBES contributions are realizable will require research studies comparing the cost and error effects of PBES estimators and estimators of conventional establishment and population sample surveys.

References

Birnbaum, Z. And Sirken, M. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. National Center for

Health Statistics. *Vital and Health Statistics,* 2(11).

Cox, G.B., Folsom, R.E., And Virage, T.G. (1987). Design Alternatives for Integrating the National Medical Expenditure Survey with the National Health Interview Survey. National Center for Health Statistics. *Vital and* Health *Statistics*, 2(101).

Judkins, D., Marker, D, And Waksberg, J. (1995). *National Health Care Survey: List Verses Network Sampling.* Unpublished report. National Center for Health Statistics.

Judkins, D., Marker, D., Waksberg, J., Botman, S., And Massey, J. (1999). National Health Interview Survey: Research for the 1995-2004 Redesign. National Center for Health Statistics. *Vital and Health Statistics,* Series 2(126). 76-80.

Leaver, S. And Valliant, R. (1995). Statistical Problems in Estimating the U.S. Consumer Price Index. In Cox, Binder, Chinnappa, Christianson, Cooledge, and Kott.(eds.) *Business Survey Methods.* John Wiley and Sons, Inc.

Massey, J.T., Moore, T.F., Parsons, V., And Tadros, W. (1991). Design and Estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics. *Vital and Health Statistics Series*, Series 2(110).

Mathiowetz, N. (1987). Linking the National Survey of Family Growth With the National Health Interview Survey: Analysis of Field Trials. National Center for Health Statistics. *Vital and Health Statistics Series*, Series 2(103).

Shimizu, I. And Sirken, M. (1998). More on Population Based Establishment Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 7-12.

Sirken, M., Shimizu, I., And Judkins, D. (1995). The Population Based Establishment Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association.* Vol. 1, 470-473.

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

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

Waksberg, J. And Northrup, D. (1985). Integration of the Sample Design of the National survey of Family Growth, Cycle IV, with the National Health Interview Survey. National Center for Health Statistics. *Vital and Health Statistics*, Series 2(96).

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