Journal of the Nigerian Association of Mathematical Physics
***Volume28, No. 1*, (*November, 2014*), pp 493 – 498**

Performance of Horvitz-Thompson Estimator In Population Based Establishment Sample Surveys Over Multiplicity Estimator

[1]Amalare A.A . and [2]Agbona A. A

[1]Lagos State Polytechnic, Department of Mathematics/Statistics,Ikorodu
[2]Federal Polytechnic, Department of Statistics, Ede , Osun State.

Abstract

In this paper, PBES (Population Based Establishment Survey) Horvitz-Thompson Estimator is discussed and considered to be an attractive design alternative to the conventional establishment sample survey and is applicable whenever free standing sampling frames are inadequate . In network sampling, the results revealed that PBES Horvitz-Thompson estimator counts each distinctnetwork only once compare to he PBES multiplicity estimator. The derivation made in respect ofunbiased PBES Horvitz-Thompson Estimator and its variance,shows to give a more robust, efficient and consistent estimator which produces good approximation results than PBES MultiplicityEstimator.

Keywords:PBES Horvitz-Thompson Estimator, Network Sampling, Establishment transactions, PBES Multiplicity Estimator, Integrated Sample Design.

1.0 Introduction

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) Horvitz-Thompson Estimatorwill serve asan 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 also plays an important roles over 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_j^{th} (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 design as two-stage sample surveys in which establishment 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[1,2], 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 surveysis a notable exception [3]. 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 frames for the CPI Pricing Survey, a sample survey of retail establishment that collects prices for a basket of consumer goods.

Several years ago, a Panel of the Committee on National Statistics, National Research of US Council[4], 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 [5]. 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 as suggested by Massey et. al [6]. The committee's suggestion initiated a PBES research program at NCHS . Judkins et. al [7] 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 et. al [8] made rough cost/error comparisons of an

Corresponding author: ***Amalare A.A.,*** E-mail: amalareasm@yahoo.com, Tel.: +2348023998403

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 close formulas for their variances. Conceptual difficulties initially encountered in this effort were overcome since it was recognized by Sirken [9], that the PBES is a population network sample survey . Applying network sampling theory , Sirken and his co-workers [10,11] obtained two versions of the unbiased PBES multiplicity estimator and derived their variances.

In this paper, the unbiased PBES Horvitz-Thompson estimator and its variance are presented. The PBES estimators are essentially extensions to multiple stage sampling under special conditions of single-stage network sampling estimators that were originally proposed by Birnbaum and Sirken [12]and described by Thompson [13] .

2.0 Notation

Let X represent the number of transactions of the E_j^{th} (j = 1,…, R) establishment. Then

$$M = \sum_{j=1}^{R} M_j = thetotalnumberoftransactionsoftheRestablishments. \qquad (1)$$

Let N_j = the number of households having transactions with E_j^{th} (j = 1,…, R) establishments N_{jl} = number of households having transactions with both E_j^{th} and E_l^{th} ($j \neq l$) establishments, and N_o = number of households not having any transactions with any establishments.

Then

$$N^* = \sum_{j=1}^{R} N_j - \sum\sum_{j\neq l} N_{jl}$$
$$= thetotalnumberofhouseholdshavingtransactionswithRestablishments. \qquad (2)$$

and

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

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

$$X_j = \sum_{k=1}^{M_j} X_{jk} = sumofthevariateoverM_jtransactionoftheE_jthestablishment. \qquad (4)$$

and

$$X = \sum_{j=1}^{R} X_j = thesumofthevariateoverMtransactionsofRestablishments. \qquad (5)$$

3.0 Methodology

3.1 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. [8] 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-stagenetworksamplepopulationsurvey. 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_j^{th} 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 . Multiplicity estimators count the M_j transactions linked to the A_j^{th} (j = 1, ..., R) network everytime households belonging to the A_j^{th} network are selected in the population survey sample. The HorvitzThompson estimator does not depend on the number of times that households belonging to the same networks are selected in the population survey. The PBES HorvitzThompson estimator counts each distinct network only once.

3.2 The PBES Horvitz-Thompson Estimator

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

$$m = \sum_{j=1}^{r} m_j \, transactions \tag{6}$$

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

$$X' = \sum_{j=1}^{R} \frac{\alpha_j}{p_j} X_j \tag{7}$$

Here, $\alpha_j$is a random variable that is equal to 1 if any ofthe n sample households belongs to the A_j^{th} network and $\alpha_j$is equal to 0 otherwise, and

$$X_j' = M_j \sum_{k=1}^{m_j} \frac{M_{jk}}{m_j} \tag{8}$$

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

$P_j = E(\alpha_j)$

= the probability of any of then sample households belongingto A_j^{th} (j = 1, …,R) network. (9)

X' is an unbiased estimate of X if every one 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_j^{th} network. (10)

If n househo1d are selected by simple random samplingwithout replacement,

$$q_j = \frac{\binom{N-N_j}{n}}{\binom{N}{n}} \tag{11}$$

If n households are selected by simple random samplingwith replacement,

$$q_j = \frac{\left(N - N_j\right)^n}{(N)^n} \tag{12}$$

There are two potential measurement problems involving the q_js (j = 1, ..., r). First, they are dependenton the N_js (j = 1, ..., r), quantities that are often difficultto ascertain in establishment surveys. Second, it wouldbe difficult to compute the q_js for most populationsurveys which, like the NHIS, are based on complex sample designs.

3.3 The Variance of the PBES Horvitz-Thompson Estimator

The variance of the Horvitz-Thompsonestimator of Xmay be written as

$$Var(X') = VarE(X'|\Omega) + E(VarX'|\Omega) \tag{13}$$

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

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

$$VarE(X'|\Omega) = Var\left(\sum_{j=1}^{R} \frac{\alpha_j X_j}{p_j}\right) = \sum_{j=1}^{R} \frac{X_j^2}{p_j^2} Var(\alpha_j) + \sum_{j=1}^{R} \sum_{l \neq j}^{R} \frac{X_j}{p_j} \frac{X_l}{p_l} Cov(\alpha_j \alpha_l) \tag{14}$$

Since α_j is a binomial random variable

$$Var(\alpha_j) = p_j - p_j^2 \tag{15}$$

and

$$Cov(\alpha_j \alpha_l) = p_{ji} - p_j p_l \tag{16}$$

Where

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

is the joint probability that any of the n samplehouseholds belongs to theA_j^{th} and the A_l^{th} networks, and

q_{jl}^* (j≠l) is the probability that the n sample househo1ds are linked to neitherA_j^{th} norA_l^{th} network.

For simple random sampling of n households withreplacement,

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

and for simple random sampling of n households withoutreplacement,

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

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

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

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

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

Where the population variance

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

Optimum allocation of m transactionsto 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

$$Var(X') = \sum_{j=1}^{R}\frac{1 - p_j}{p_j}X_j^2 + \sum_{j=1}^{R}\sum_{l\neq j}\frac{p_{jl} - p_j p_l}{p_j p_l}X_j X_l + \sum_{j=1}^{R}\frac{M_j^2}{p_j}\frac{M_j - m_j}{m_j M_j}\sigma^2\left(X_j\right). \qquad (24)$$

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.

All unbiased PBES estimators, whether the Horvitz-Thompson estimator proposed in this paper or the PBES multiplicity estimatorsproposed by Sirken, Shimizu, andJudkins [10] and Shimizu and Sirken [11] depend onmultiplicity parameters to adjust for variations in theselection probabilities of the establishments reported inthe population sample survey. However, multiplicity andHorvitz-Thompson estimators differ in the waysmultiplicities are defined and in likelihoodof successfullycollecting this information in the follow-on survey withthe establishments that were reported in the populationsurvey.

The N_js and M_js(j = 1, ..., r) respectively are the multiplicities neededthe PBES Horvitz-Thompson estimator and the PBES multiplicity estimators, where N_j is the number of households having transactions with the E_j^{th} establishment, and M_j is the total number of transactionsof the E_j^{th}

establishment. The N_js are unlikely to bereadily available except at establishments, such as healthmaintenance organizations, utility companies, and home owner insurance companies, for which households are the transactional units.

On the other hand, the M_js 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.

4.0 Conclusion

The PBES Horvitz-Thompson Estimator is discussed in network sampling method . It is shown that PBES Horvitz-Thompson Estimator usually counts each distinct network only once compare to the PBES multiplicity estimator since the PBES multiplicity estimator depend on the number of times the households are selected in the population survey . The derivation made in respect of unbiased PBES Horvitz-Thompson Estimator and its variance,shows to give a more robust, efficient and consistent estimator which produces good approximation results than PBES Multiplicity Estimator .
However, It is affirmed that the PBES Horvitz-Thompson Estimator 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.

References

[1] Mathiowetz, N. (1987). ‘ Linking the National Survey of Family Growth with the National HealthInterview Survey:Analysis of Field Trials ‘. National Center for Health Statistics . Vital and Health Statistics , Series 2 (103) , 15 – 19.

[2] 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), 32 – 36 .

[3] 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, New York .

[4] Wunderlich, G.S. (ed)(1992). ‘Towards a National Health Care Survey ‘: A Data System for the 21stCentury. National Academy Press, Washington, D.C.

[5] 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), 43 – 49.

[6] 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 2(110) , 71 – 75 .

[7] Judkins, D., Marker, D, and Waksberg, J. (1995). ‘ National Health Care Survey: List Verses Network Sampling‘. John Wiley and Sons, New York .

[8] 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 .

[9] Sirken, M. (1970).’Household Surveys with Multiplicity‘. Journal of the American Statistical Association, 65, 257-266

[10] 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 .

[11] Shimizu, I. and Sirken, M. (1998). ‘ More on population Based Establishment Surveys. Proceedings of the Section on Survey Research Methods’, American Statistical Association, Vol.2, 7-12.

[12] 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) , 67 - 71 .

[13] Thompson, S. (1992).’ Sampling ‘. John Wiley and sons, New York