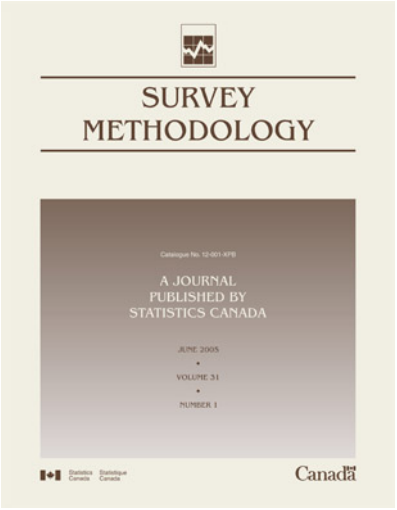
Catalogue no. 12-001-XIE

Survey Methodology

December 2002



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2002

Published by authority of the Minister responsible for Statistics Canada

January 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 28, No. 2, pp. 183-190
Statistics Canada, Catalogue No. 12-001

Design Effects of Sampling Frames in Establishments Survey

Monroe G. Sirken [1]

Abstract

When stand-alone sampling frames that list all establishments and their measures of size are available, establishment surveys typically use the Hansen–Hurwitz (HH) pps estimator to estimate the volume of transactions that establishments have with populations. This paper proposes the network sampling (NS) version of the HH estimator as a potential competitor of the HH estimator. The NS estimator depends on the population survey-generated establishment frame that lists households and their selection probabilities in a population sample survey, and the number of transactions, if any, of each household with each establishment. A statistical model is developed in this paper to compare the efficiencies of the HH and NS estimators in single-stage and two-stage establishment sample surveys assuming the stand-alone sampling frame and the population survey-generated frame are flawless in coverage and size measures.

Key Words: Stand-alone establishment frames; Population survey-generated establishment frames; Hansen-Hurwitz estimator; Network sampling estimator.

1. Introduction

Listings of establishments that have transactions with households in population sample surveys serve as sampling frames of establishment surveys whenever the transactions reported by households in the population surveys are matched with the records of their establishments. For example, the listings of establishments that have transactions with households in the National Medical Expenditure Panel Survey (MEPS), a national population sample survey, serve as sampling frames for medical provider surveys that supplement and verify the medical expenditures of the transactions reported by MEPS household respondents (Cohen 1998). However, listings of establishments that have transactions with households in population sample surveys rarely serve as frames of establishment surveys that collect information about the transactions that establishments have with all households. The Current Price Index (CPI) produced by the Bureau of Labor Statistics is a notable and rare exception of a Federal establishment survey that depends on a population survey-generated sampling frame. The CPI Pricing Survey, a national retail establishment survey, that collects prices for a basket of consumer goods purchased by all customers, uses as its sampling frame the listings of retail establishments that have transactions with households in the CPI Continuing Point of Purchase Survey. (Leaver and Valliant 1995).

After reviewing plans of the National Center for Health Statistics (NCHS) to restructure its family of independent national surveys of health providers (hospitals, physicians, clinics, *etc*.), a Panel of the Committee on National Statistics proposed (Wunderlich 1992) using listings of health care providers reported by households in the National Health Interview Survey (NHIS), an ongoing national household sample survey (Massey, Moore, Parsons and Tadros 1991) as the sampling frames for national surveys of health care providers. The Committee thought that, especially in the current environment of rapid changes in listings of health care providers due to rapid changes in the nation's health care delivery system, the NHIS-generated health care provider frames would be more accurate and easier and less expensive to construct and maintain than the free-standing health care provider frames currently in use. Soon after the Panel report was issued, NCHS initiated a research project on population survey-generated sampling frames that is briefly summarized below.

Initially, the research focused almost exclusively on the statistical properties of NHIS-generated frames of health care providers. Judkins, Berk, Edwards, Mohr, Stewart and Waksberg (1995) studied the quality of the free-standing health provider frames currently in use or of potential use, and discussed the kinds of medical providers for which NHIS-generated frames would seem to have the greatest potential. Subsequently, Judkins, Marker, Waksberg, Botman and Massey (1999) made rough comparisons of the efficiencies of dental surveys using the NHIS-generated sampling frame and using the free-standing frame, and concluded that NHIS-generated health care provider frames deserve serious consideration whenever reasonably complete free-standing health care provider frames with reasonably good size measures are unavailable.

In recent years, the research has focused on the statistical properties of estimators that depend on population-generated sampling frames and has become more theoretically focused than formerly. The conceptual difficulties initially encountered in developing unbiased estimators for the population survey-generated frame because the same establishments have transactions with multiple households were overcome by applying network sampling theory. (Sirken

1. Monroe G. Sirken, Senior Research Scientist, National Center for Health Statistics, U.S.A.

1997; Thompson 1992). Sirken, Shimizu and Judkins (1995) developed the network sampling version of the HH estimator, referred to in this paper as the NS estimator, and Sirken and Shimizu (1999) developed the network sampling version of the Horwitz–Thompson (HT) estimator. This paper develops a statistical error model that compares the efficiencies of the NS estimator that depends on the population survey-generated frame, and the HH estimator that depends on the free-standing frame. The error model assumes both frames are flawless in establishment coverage and size measures and have equivalent construction and maintenance costs. Though the model assumes a srs design for the population survey that generates population survey-generated sampling frame, the model can be applied to other kinds of population survey designs that are not considered in this paper.

This paper is organized as follows. Notation follows in section 2. Section 3.1 and section 3.2 respectively present the pps self-weighted HH estimator and variance of the two-stage establishment sample survey that depends on the free-standing sampling frame, and the NS estimator and variance of a two-stage establishment survey that depends on the population survey-generated frame. The error model is developed in sections 4.1–4.4. The difference between two-stage HH and NS variances of equivalent expected sample sizes is developed in section 4.1. In section 4.2, the first stage variance component of the two-stage NS estimator is split into variance components representing effects of households with and without transactions, and section 4.3 shows the design effects of the NS estimator in single stage sampling. Second stage variance components of the NS and HH estimators are compared in section 4.4. In the concluding section 5, the error model's major findings comparing efficiencies of HH and NS estimators in single-stage and two-stage establishment surveys are briefly summarized, and limitations of the model are briefly discussed. The appendix presents the proof of a statistical statement appearing in section 4.2.

2. Notation

Let $N_j =$ the number of households having transactions with establishment $j(j = 1, 2, ..., R)$, $N_o =$ the number of households not having transactions with any establishments, and $N^* =$ the number of distinct households having transactions with R establishments. Then, $N = N^* + N_o =$ the total number of households.

Let $M_{ij} =$ the number of transactions of establishment $j(j = 1, 2, ..., R)$ with household $i(i = 1, 2, ..., N)$, where $M_{ij} \geq 0$ when establishment j has transactions with household i, and $M_{ij} = 0$ when establishment j and household i do not have transactions. Then, $M_j = \sum_{i=1}^{N} M_{ij} =$ the number of transactions of establishment j with N households, and $M = \sum_{j=1}^{R} M_j =$ the number of transactions of M establishments with N households, and $\overline{M} = M / N$ the average number of transactions per household.

Let X_{jk} denote the value of the x-variate for transaction $k(k = 1, ..., M_j)$ of establishment $j(j = 1, 2, ..., R)$. Then, $X_j = \sum_{k=1}^{M_j} X_{jk} =$ the sum of the x-variate over the M_j transactions of establishment j, and $X = \sum_{j=1}^{R} X_j =$ sum of the x-variate over the M transactions of R establishments. Let $\overline{X}_j = X_j / M_j =$ the average value of the x-variate over the M_j transactions of establishment j, and $\overline{X} = X / M =$ the average value of the x-variate over M transactions.

3. Estimators and Variances

3.1 The HH Estimator and Variance

Consider a two-stage self weighted establishment sample survey using a free-standing establishment sampling frame that lists all R establishments and their measures of size, $M_j (j = 1, 2, ..., R)$. Establishments are the primary sampling units (psu's), and transactions are the secondary sampling units. A sample of r establishments is selected with pps with replacement from the free-standing frame, and a sample of size $t_{\mathrm{HH}} < \min(M_1, ..., M_j, ..., M_R)$ transactions each, where t_{HH} is a positive integer, is independently selected by simple random sampling without replacement for each sample establishment $j(j = 1, 2, ..., r)$.

The unbiased self-weighted pps HH estimator of X is

$$X'_{\mathrm{HH}} = \frac{M}{r} \sum_{j=1}^{r} \overline{X}'_j \qquad (1)$$

where $\overline{X}'_j = \sum_{k=1}^{t_{\mathrm{HH}}} X_{ij} / t_{\mathrm{HH}}$ is the unbiased estimate of $\overline{X}_j = X_j / M_j (j = 1, 2, ..., R)$. Because establishments are selected with replacement, the HH estimator counts $\overline{X}_j$ as many times as establishment j is selected in the sample.

The variance of the X'_{HH} is (Thompson 1992)

$$\mathrm{Var}(X'_{\mathrm{HH}}) = \frac{M^2}{r} \sigma^2_{\mathrm{HH1}} + \frac{M}{r t_{\mathrm{HH}}} \sum_{j=1}^{R} (M_j - t_{\mathrm{HH}}) \sigma^2_j \qquad (2)$$

where the first and second terms respectively on the right side of (2) are the first and second stage variance components, and

$$\sigma^2_{\mathrm{HH1}} = \frac{1}{M} \sum_{j=1}^{R} M_j (\overline{X}_j - X / M)^2 \qquad (3)$$

is the between establishment population variance, and

$$\sigma^2_j = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (X_{jk} - X_j / M_j)^2 \qquad (4)$$

is the within establishment population variance of establishment j.

3.2 The NS Estimator and Variance

Consider a two-stage establishment sample survey that depends on a population survey-generated frame. The frame lists n sample households $H'_i (i = 1, 2, ..., n)$ that were

enumerated in a population sample survey. For each listed household H_i', the frame provides π_i, its selection probability in the household survey, and M_{ij}, the number of its transactions with each distinct establishment $j(j=1, 2, ..., R)$ (the M_{ij}'s are reported by household respondents in the population sample survey).

Each of the n listed households in the population survey-generated frame represents a cluster of establishments ranging in size from 0 to R establishments with whom the household has transactions. The n clusters of establishments are the primary sampling units, and the $M_j(j=1, 2, ..., r)$ transactions of the r sampled establishments are secondary sampling units. The transaction sample for establishment $j(j=1, 2, ..., R)$ is selected as follows: a srs sample of size $t_{\text{NS}} M_{ij} < \text{Min}(M_1, M_2, ..., M_r)$ transactions is independently selected without replacement for each sample household $H_i'(i=1, 2, ..., n)$, where t_{NS} is a positive integer. The transaction sample size of establishment $j(j=1, 2, ..., R)$ is equal to $t_{\text{NS}} \sum_{i=1}^{n} M_{ij}$, and the total transaction sample size is equal to τt_{NS}, where $\tau = \sum_{i=1}^{n} \sum_{j \in A_i} M_{ij} =$ the sum of the transactions over n sample households is a random variable.

The NS estimator of X is

$$X'_{\text{NS}} = \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$$

where A_i is the cluster of distinct establishments that have transactions with sample household H_i, and

$$\bar{X}'_j(i) = \sum_{k=1}^{t_{\text{NS}} M_{ij}} X_{jk} \Big/ (t_{\text{NS}} M_{ij})$$

is an unbiased estimate $\bar{X}_j$ for a sample of $t_{\text{NS}} M_{ij}$ transactions of establishment j. Because households are selected with replacement, the NS estimator counts the quantity $\sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$ every time household $H_i(i=1, 2, ..., n)$ is selected in the sample, and because the same establishment has transactions with multiple households, the NS estimator counts the quantity $M_{ij} \bar{X}'_j(i)$ every time a sample household $i(i=1, 2, ..., n)$ contains establishment j.

Assuming a srs design in the population survey, $\pi_i = n/N$, and the network sampling estimator is

$$X'_{\text{NS}} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j \in A_i} M_{ij} \bar{X}'_j(i). \tag{5}$$

The NS estimator is an unbiased estimator of X.

$$E(X'_{\text{NS}}) = \sum_{i=1}^{n} E \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \sum_{i=1}^{N} \sum_{j \in A_i} M_{ij} \bar{X}'_j$$

$$= \sum_{i=1}^{R} M_j \bar{X}_j = \sum_{J=1}^{R} X_j = X.$$

The NS estimator in (5) is self-weighted because we have assumed that the n households are selected by srs. It would be a self-weighted estimator whenever the sample design of the population sample survey that generates the establishment sampling frame is self-weighted. When $N = N^* = M$, implying that N^* households each has a single transaction, and $N_0 = N - N^*$ households are without transactions, and when $n = r$ and $t_{\text{NS}} = t_{\text{HH}}$, the HH and NS estimators are equivalent.

$$X'_{\text{NS}} = \frac{N}{n} \sum_{i=1}^{M} \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \frac{N}{n} \sum_{i=1}^{N} \sum_{j \in A_i} \bar{X}'_j$$

$$= \frac{M}{r} \sum_{j=1}^{R} \bar{X}'_j = X_{\text{HH}}. \tag{6}$$

The variance of the NS estimator (5), under srs sampling with replacement of n households and independent selections of $t_{\text{NS}} M_{ij}$ transaction by srs without replacement for each establishment j linked to household H_i, is (Sirken *et al*. 1995)

$$\text{Var}(X'_{\text{NS}}) = \frac{N^2}{n} \sigma^2_{\text{NS1}} + \frac{N}{n t_{\text{NS}}} \sum_{i=1}^{N} \sum_{j=1}^{R} M_{ij} \frac{M_j - t_{\text{NS}} M_{ij}}{M_j} \sigma_j^2 \tag{7}$$

where the first and second terms respectively on the right side of (7) are the first and second stage variance components, and

$$\sigma^2_{\text{NS1}} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j \in A_i} M_{ij} \bar{X}'_j - X/N \right)^2 \tag{8}$$

is the population variance between households, and σ_j^2, the population variance within establishment j as defined in (4). An unbiased estimate of NS variance is

$$\text{Var}(X'_{\text{NS}}) = \frac{N^2}{n(n-1)} \sum_{i=1}^{n} \left[\sum_{j \in A_i} M_{ij} \bar{X}'_j(i) - \bar{X}' \right]^2 \tag{9}$$

where $\bar{X}' = X'/N$.

4. The Error Model

4.1 HH and NS Variances of Equivalent Expected Sample Size

Subtracting (2) from (7), the difference between the variances of the HH and NS estimators of X is

$$\mathrm{Var}(X'_{\mathrm{NS}}) - \mathrm{Var}(X'_{\mathrm{HH}}) = \left[\frac{N^2}{n}\sigma^2_{\mathrm{NS1}} - \frac{M^2}{r}\sigma^2_{\mathrm{HH1}}\right] + \left[\frac{N}{nt_{\mathrm{NS}}}\sum_{i=1}^{N}\sum_{j=1}^{R} M_{ij}\frac{M_j - t_{\mathrm{NS}}M_{ij}}{M_j}\sigma^2_j - \frac{M}{rt_{\mathrm{HH}}}\sum_{j=1}^{R}(M_j - t_{\mathrm{HH}})\sigma^2_j\right] \quad (10)$$

where the first and second set of bracketed terms respectively on the right side of (10) represent the differences between the primary and secondary variance components of the HH and NS estimators of *X*.

Let $m_{\mathrm{NS}} = \tau t_{\mathrm{NS}} =$ the size of the transaction sample in the establishment survey using the population survey-generated frame, where t_{NS}, a positive integer, is the size of the transaction sample selected per transaction of the *n* sample households, and $\tau = \sum_{i=1}^{n}\sum_{j\in A_i} M_{ij} =$ sum of the transactions of *n* sample households.

Clearly, τ is a random variable and its expected value conditional over all samples of *n* households is $E(\tau|n) = n\overline{M}$ where $\overline{M} = M/N =$ average household transaction size. It follows that $E(m_{\mathrm{NS}}|n) = t_{\mathrm{NS}}E(\tau|n) = n\,\overline{M}t_{\mathrm{NS}}$ is the expected transaction sample size of the NS estimator conditional over all samples of *n* households.

Let $m_{\mathrm{HH}} = rt_{\mathrm{HH}} =$ the size of the transaction sample in the establishment survey using the stand-alone frame, where $r =$ the establishment sample size, and $t_{\mathrm{HH}} =$ the transaction sample size per selected establishment. Let $r = E(\tau|n) = n\,\overline{M}$ and let $t_{\mathrm{HH}} = t_{\mathrm{NS}} = t$, and it follows the expected transaction sample sizes of the NS and HH estimators conditional over all samples of *n* households are equivalent, namely, $E(m_{\mathrm{HH}}|n) = tE(\tau|n) = nt\overline{M} = E(m_{\mathrm{NS}}|n)$.

Calibrating the establishment and transaction sample sizes in this manner assures that HH and the NS establishment surveys are conducted under roughly the same fiscal constraints if per establishment and per transaction field costs are about the same in both surveys. It is noteworthy, however, that this cost equation does not take into account the differences in costs between constructing and maintaining stand-alone establishment frames and population survey-generated establishment frames.

Substituting $r = n\overline{M}$, $t_{\mathrm{HH}} = t_{\mathrm{NS}} = t$, and $\mathrm{M} = \mathrm{N}\,\overline{M}$ in formula (9), the difference between the NS and HH variances of equivalent expected establishment and transaction sample size conditional over all samples of *n* households is

$$\mathrm{Var}(\mathrm{X}'_{\mathrm{NS}}) - \mathrm{Var}(\mathrm{X}'_{\mathrm{HH}}) = \frac{N^2}{n}\left[\sigma^2_{\mathrm{NS1}} - \overline{M}\sigma^2_{\mathrm{HH1}}\right] - \frac{N}{nt}\sum_{j=1}^{R}\sigma^2_j\left[(M_j - t) - \sum_{i=1}^{N}\frac{M_{ij}(M_j - M_{ij})}{M_j}\right]. \quad (11)$$

The first term and second terms respectively on the right side of (11) represent the difference between the first stage and second stage variance components of the NS and HH estimators of equivalent expected sample sizes conditional over all samples of *n* households.

4.2 Decomposition of the Single Stage NS Population Variance

Typically, some households do not have transactions with any establishments, and the percentage varies by type of establishment. For example, medical care utilization by families in the United States varies greatly by type of health care provider (Dicker and Sunshine 1987). During a 12 month period, 70 percent of families were not admitted to hospitals, 7 percent did not have ambulatory physician visits, and 28 percent did not have dental visits.

Let

$P = \dfrac{N^*}{N} =$ fraction of N households with one or more transactions, and

$P_0 = 1 - P\dfrac{N_0}{N} =$ fraction of N households without any transactions.

We demonstrate in the Appendix that the single stage population variance of the NS estimator of *X*, when expressed as a function of *P*, decomposes into 2 parts

$$\sigma^2_{\mathrm{NS1}}(P) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\in A_i} M_{ij}\,\overline{X}_j - \frac{X}{N}\right)^2 = P\sigma^2_{\mathrm{NS1}^*} + \sigma^2(P)E^2_{\mathrm{NS1}^*}\ 0 < P \le 1 \quad (12)$$

where

$$\sigma^2_{\mathrm{NS1}} = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\overline{X}_j - \frac{X}{N^*}\right)^2 \quad (13)$$

is the single stage population variance of the *x*-variate over the truncated population of N^* households with one or more transactions,

$$E^2_{\mathrm{NS1}^*} = \left(\frac{X}{N^*}\right)^2 = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\,\bar{X}_j\right)^2 - \sigma^2_{\mathrm{NS1}^*} \qquad (14)$$

is the expected value squared of the *x*-variate over the truncated population of N^* households and

$$\sigma^2(P) = P(1-P) \qquad (15)$$

is the variance of the binomial variable *P*. For fixed *M*, the function $\sigma^2_{\mathrm{NS1}}(P|M)$ is maximum when

$$P = P_{\max} = \frac{1}{2}\left[\left(\sigma^2_{\mathrm{NS1}^*}/E^2_{\mathrm{NS1}^*}\right)+1\right] \le 1.$$

If $\sigma^2_{\mathrm{NS1}^*} \ge E^2_{\mathrm{NS1}^*}$, $P_{\max} = 1$ and if $\sigma^2_{\mathrm{NS1}} < E^2_{\mathrm{NS1}^*}$, $1/2 < P_{\max} < 1$.

When $P = 1$, $\sigma^2(P=1) = 0$ and therefore $\sigma^2_{\mathrm{NS1}}(P=1) = \sigma^2_{\mathrm{NS1}^*}$. If $P = \bar{M} = (M/N) = 1$, implying that each of *N* households has a single transaction,

$$\sigma^2_{\mathrm{NS1}}(P = \bar{M} = 1) = \sigma^2_{\mathrm{NS1}^*}(N^* = M) = \sigma^2_{\mathrm{HH1}} \qquad (16)$$

because

$$\sigma^2_{\mathrm{NS1}^*}(N^* = M) = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\bar{X}_j - \frac{X}{N^*}\right)^2$$

$$= \frac{1}{M}\sum_{j=1}^{R} M_j\left(\bar{X}_j - \frac{X}{M}\right)^2 = \sigma^2_{\mathrm{HH1}}, \qquad (17)$$

and, $\sigma^2(P=1) = 0$. In other words when $P = \bar{M} = 1$, implying each of the *N* households has a single transaction, the variance of the NS1 estimator which would then depend on a srs of transactions with replacement is equivalent to the variance of the HH1 estimator that depends on a pps cluster sample of equivalent sample size selected with replacement.

4.3 Design Effects in Single Stage Sampling

Let

$X'_{\mathrm{NS1}} = \frac{N}{n}\sum_{i=1}^{N}\sum_{j\in A_i} M_{ij}\bar{X}_j =$ the unbiased NS estimator of *X*

in single stage sampling, and

$X'_{\mathrm{HH1}} = \frac{M}{R_{\mathrm{HH}}}\sum_{j=1}^{r_{\mathrm{HH}}}\bar{X}_j =$ the unbiased HH estimator of *X* in

single stage sampling.

Define the single stage sampling total design effect of the NS1 estimator as the ratio of the variances of the NS1 and HH1 estimators of equivalent sample size conditional over all samples of *n* households.

$$\lambda(P) = \frac{\mathrm{Var}(X'_{\mathrm{NS1}})}{\mathrm{Var}(X'_{\mathrm{HH1}})} = \frac{\sigma^2_{\mathrm{NS1}}(P)}{\bar{M}\sigma^2_{\mathrm{HH1}}} \qquad (18)$$

where $\lambda(P) < 1$ indicates that the NS1 estimator is more efficient than the HH1 estimator, and $\lambda(P) > 1$ indicates that the HH1 estimator is more efficient than the NS1 estimator.

We noted in (12) and (15) that $\sigma^2_{\mathrm{NS1}}(\mathrm{P}) = \mathrm{P}\,\sigma^2_{\mathrm{NS1}^*} + \mathrm{P}(1-P)(X/N^*)^2$, and in (16) that $\sigma^2_{\mathrm{HH1}} = \sigma^2_{\mathrm{NS1}^*}(N^* = M)$. Making these substitutions in (18), the total design effect becomes

$$\lambda(P) = \mathrm{deft}^2_{\mathrm{NS1}} + (1-P)\;Z_{\mathrm{NS1}},\; 0 < P \le 1 \qquad (19)$$

where

$$Z_{\mathrm{NS1}} = \frac{P(X/N^*)^2}{\bar{M}\,\sigma^2_{\mathrm{NS1}^*}(N^* = M)} \qquad (20)$$

is the effect due to the N_o households without transactions, and

$$\mathrm{deft}^2_{\mathrm{NS1}} = \left[\frac{\mathrm{P}\,\sigma^2_{\mathrm{NS1}^*}}{\bar{M}\,\sigma^2_{\mathrm{HH1}}}\right] = \left[\frac{\mathrm{P}\,\sigma^{2^*}_{\mathrm{NS1}}}{\bar{M}\,\sigma^2_{\mathrm{NS1}^*}(N^* = M)}\right] \qquad (21)$$

is effect due to the N^* households with transactions. In other words, $\mathrm{deft}^2_{\mathrm{NS1}}$ is the design effect of network sampling a population of N^* household clusters containing one or more transactions, with equal probability and replacement, compared to *network sampling* a population of *M* transactions, of equivalent expected sample size, by srs and replacement. [The reader is referred to Kish (1982) for the definition of deft^2].

The total design effect in (19) depends on $\mathrm{deft}^2_{\mathrm{NS}}$ and $\mathrm{Z}_{\mathrm{NS1}}$ and, P, and the values of these parameters, as well as relationships between them, are likely to vary considerably between surveys, and between variables and population domains in the same surveys. Though, in theory, the NS1 estimator could be more efficient than HH1 estimator, in reality that outcome seems highly unlikely because cluster sampling is typically less efficient than srs. A necessary condition for the NS1 estimator to be as efficient or more efficient than the HH1 estimator is that $\mathrm{deft}^2_{\mathrm{NS1}} \le 1-(1-\mathrm{P})\mathrm{Z}_{\mathrm{NS1}}$, and this condition is unlikely to be met particularly if *P* is small, and if the within household transaction clustering is mostly due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments.

4.4 Comparing Efficiencies in Two-stage Sampling

In two stage sampling, the difference between the HH and NS second stage variance components for equivalent expected sample size of $nt\bar{M}$ transactions conditional over

all samples of n households, the second term on the right side of equation (11), reduces to

$$\frac{N}{nt}\sum_{j=1}^{R}\sigma_j^2\left\{(M_j - t) - \sum_{j=1}^{N}\frac{M_{ij}(M_j - tM_{ij})}{M_j}\right\} = \frac{N}{n}\sum_{j=1}^{R}\frac{\rho_j}{M_j}\sigma_j^2 \qquad (22)$$

where $\rho_j / M_j = 1/M_j \sum_{i=1}^{N} M_{ij}(M_{ij} - 1)$ is the difference between the HH and NS second stage finite population corrections for establishment j. If none of the N households have multiple transactions with establishment j, the HH and NS second stage variances of establishment j are equivalent and $\rho_j = 0$. Otherwise, $\rho_j > 0$ and second stage variance for establishment j is larger for the HH than the NS estimator. The value of ρ_j is maximum when establishment j has M_j transactions with a single household.

The second stage variance components of the HH and NS estimators are equivalent $\sum_{j=1}^{R}\rho_j = 0$, when, that is, none of the H households have multiple transactions with any of the R establishments. Of course, second stage variances are equivalent if transactions are selected with replacement or the within establishment variances, $\sigma_j^2 = 0 (j = 1, 2, ..., R)$. Except for these contingencies, however, the second stage variance is always larger for the HH estimator than for the NS estimator, and the magnitude of the difference depends on the extent of within household clustering of transactions with the same establishments, and the magnitudes of the within establishment variances.

If none of the N^* households have multiple transactions with the same establishments, the difference between the variances of the HH and NS estimators are equivalent in single stage and two stage establishment sample surveys. Otherwise, the difference between HH and NS variances is less in two stage than in single stage establishment sample surveys because whenever households have multiple transactions with the same establishments the second stage variance is greater for the HH estimator than for the NS estimator.

5. Summary and Concluding Remarks

The error model presented in this paper compares efficiencies of two estimators of the volume of transactions between establishments and populations in single-stage and two-stage establishment sample surveys. The Hansen-Hurwitz (HH) estimator depends on a stand-alone sampling frame that lists every establishment and the volume of its transactions with all households during a specified calendar period. The network sampling (NS) estimator depends on a population survey-generated frame that lists the households and their selection probabilities in a population sample survey, and for each household, lists the number of its transactions with each distinct establishment during the specified calendar period.

Also, the NS and HH estimators depend on different establishment survey sample designs. In single-stage sampling, the HH estimator depends on a design in which establishments are the selection units and they are selected with pps with replacement, and the NS estimator depends on a design in which households are the selection units and they are selected with their selection probabilities in the population survey, which the error model assumes is srs with replacement. In two-stage sampling, transactions are the second stage sampling units of the HH and NS estimators. The HH estimator depends on fixed-size transaction samples that are selected by srs independently without replacement. The NS estimator depends on transaction sample sizes that are proportional to the number of transactions of each household with each establishment, and are selected independently by srs without replacement.

The NS and HH estimators are equally efficient, if and only if, every household in the entire population has one and only one transaction. Otherwise, neither the NS or the HH estimator is necessarily more efficient than the other. Nevertheless, it seems likely that the HH estimator will be more efficient than the NS estimator in single-stage establishment survey sampling, and perhaps substantially more efficient especially when large fractions of households do not have any transactions, and/or when the within household clustering of transactions among households with transactions is principally due to households having multiple transactions with the same establishments rather than households having transactions with multiple establishments. In two-stage sampling, the outcome is not as transparent as in single stage sampling because the second stage variance component is larger for the HH estimator than the NS estimator by an amount that depends on the extensiveness of within household clustering of transactions with the same establishments.

Arguably the foremost limitation of the error model presented in this paper is the presumption that the stand-alone and population survey-generated sampling frames are flawless in coverage and size measures. However, comparative costs of constructing and maintaining good quality stand-alone and population-generated establishment sampling frames are likely to vary greatly from survey to survey. Though the model seek to equalize the establishment survey costs based on each kind of sampling frames it ignores the differential costs of constructing and maintaining each kinds of frame.

Even in the absence of empirical data about the comparative costs of constructing and maintaining the frames, it is fair to say that the population survey-generated frame should be seriously considered as a potential design alternative whenever constructing and maintaining good quality stand-alone frames would be infeasible or exorbitantly expensive or time consuming, and/or when constructing and maintaining good quality population survey-generated

establishment sampling frames would be relatively inexpensive. For example, the population survey-generated frame would be a particularly attractive as a potential design alternative to the stand-alone frame when the stand-alone frame would be difficult to construct and maintain because it was undergoing rapid changing due to births, deaths, and establishment mergers, and the population survey-generated frame costs would be relatively small either because it could be constructed and maintained as a by-product of an ongoing population sample survey (Wunderlich 1992) and/ or as a by-product of an ongoing program of matching transactions of households enumerated in a population survey with their establishment records (Cohen 1998).

Another limitation of the model is the unrealistic assumption that the population survey that generates the establishment sampling frame is based on a single stage sample design in which households are selected with equal probabilities and with replacement. In fact, population surveys are virtually always based on multistage sample designs in which households are selected without replacement in the final sampling stage. Typically, the srs assumption tends to significantly understate the variance of the NS estimator, and therefore would have the effect of exaggerating the relative efficiency of the NS estimator compared to the HH estimator. On the other hand, the household sampling with replacement assumption would have the opposite effects, but would be modest (Sirken 2001) compared to the srs assumtion. The error model can be applied, however, to the other population survey sample designs that are not considered in this paper.

The error model presented in this paper identifies the critical parameters that determine the relative efficiency of establishment survey estimators depending on stand-alone and population survey-generated sampling frames. Values of these parameters will vary greatly between surveys and between variables and population domains in the same surveys. Unfortunately, empirical data are currently unavailable, and they are sorely needed to estimate the model's parameters under a broad range of survey conditions. Hopefully, this paper will stimulate interest in conducting establishment surveys that depend on population survey-generated establishment sampling frames, and will lead to improvements in designing establishment surveys that estimate the volume of transactions between establishments and populations.

Acknowledgments

I thank the 2 referees and in particular an Assistant Editor for very helpful comments. The views expressed in this paper are solely those of the author and do not necessarily represent the official views or positions of the National Center for Health Statistics.

Appendix

When expressed as a function of P, the fraction of households with one or more transactions, the single stage population variance of the network sampling (NS) estimator of X

$$\sigma^2_{\mathrm{NS1}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\in A_i} M_{ij}\bar{X}_j - \frac{X}{N}\right)^2$$

decomposes into 2 parts

$$\sigma^2_{\mathrm{NS1}^*}(P) = P\sigma^2_{\mathrm{NS1}^*} + \sigma^2(P)E^2_{\mathrm{NS1}^*} \quad 0 < P \le 1$$

where

$$P = \frac{N^*}{N},$$

$$\sigma^2_{\mathrm{NS1}^*} = \frac{1}{N^*}\sum_{i=1}^{N^*}\left(\sum_{j\in A_i} M_{ij}\,\bar{X}_j - \frac{X}{N^*}\right)^2$$

is the truncated single stage population variance of the NS estimator exclusive of the $N_0 = N - N^*$ households without transactions with establishments,

$$\sigma^2(P) = P(1-P)$$

is the variance of the binomial variable P, and

$$E^2_{\mathrm{NS1}^*} = (X/N^*)^2$$

is the expected value squared of the x-variate distributed over N^* households.

Proof

$$\begin{aligned}\sigma^2_{\mathrm{NS1}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j\in A_i} M_{ij}\bar{X}_j - \frac{X}{n}\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N^*}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_j - \frac{X}{N^*}\right)^2 + \frac{N_0}{N}\left(\frac{X}{N}\right)^2. \quad \text{(A.1)}\end{aligned}$$

Add and subtract X/N^* to the first term on the right side of (A.1).

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N^*}\left(\sum_{j=1}^{R} M_{ij}\bar{X}_j - \frac{X}{n}\right)^2 \\ &= \frac{P}{N^*}\sum_{i=1}^{N^*}\sum_{j\in A_i}\left(M_{ij}\bar{X}_j - \frac{X}{N^*}\right)^2 + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2 \\ &= P\sigma^2_{\mathrm{NS1}^*}(P) + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2. \quad \text{(A.2)}\end{aligned}$$

Substitute (A.2) for the first term on the right side of (A.1).

$$\sigma^2_{NS1}(P) = P\sigma^2_{NS1^*} + P\left(\frac{X}{N^*} - \frac{X}{N}\right)^2 + (1-P)\left(\frac{X}{N}\right)^2$$
$$= P\sigma^2_{NS1^*}(P) + \sigma^2(P)E^2_{NS1^*} \qquad \text{(A.3)}$$

where

$$\sigma^2(P) = P(1-P), \text{ and } E^2_{NS1^*} = \left(\frac{X}{N^*}\right)^2.$$

References

Cohen, S.B. (1998). Sample design of the 1996 medical expenditure panel survey medical provider component. *Journal of Economic and Social Measurement*, 24, 25-53.

Dicker, M., and Sunshine, J.H. (1987). Family use of health care, United States, 1980. *National Health Care Utilization and Expenditure Survey*, Report No. 10. DHHS Pub. 87-20210.

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

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*, Washington, DC: Government Printing Office, Series 2. 126, 76-89.

Kish, L. (1982). Design effect. *Encyclopedia of the Statistical Sciences*. John Wiley & Sons, Inc. 2, 347-348.

Leaver, S., and Valliant, R. (1995). Statistical problems in estimating the U.S. consumer price index. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: John Wiley & Sons, Inc.

Massey, L.T., Moore, T.F., Parsons, V. and Tadro, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2, 110.

Sirken, M., and Shimizu, I. (1999). Population based establishment surveys: The Horvitz-Thompson estimator. *Survey Methodology*, 25, 187-91.

Sirken, M., Shimizu, I. and Judkins, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 1, 470-473.

Sirken, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics*, John Wiley & Sons, Inc. 4, 2977-2986.

Sirken, M.G. (2001). The Hansen-Hurwitz estimator revisited: PPS sampling without replacement. *Proceedings of the Section on Survey Research Methods,* American Statistical Association. In print.

Thompson, S. (1992). *Sampling*. New York: John Wiley & Sons, Inc. 117-118.

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