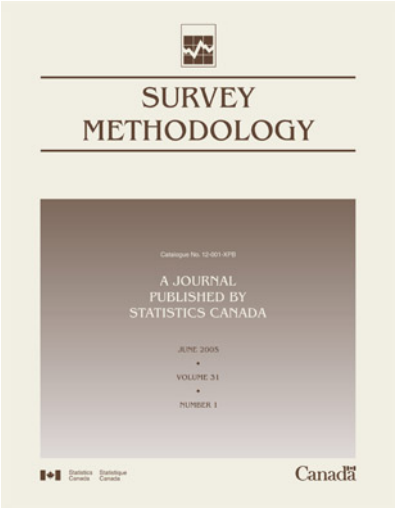
Catalogue no. 12-001-XIE

Survey Methodology

December 2003



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 2003

Published by authority of the Minister responsible for Statistics Canada

February 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. 29, No. 2, pp. 189-196
Statistics Canada, Catalogue No. 12-001

The High Entropy Variance of the Horvitz-Thompson Estimator

K.R.W. Brewer and Martin E. Donadio [1]

Abstract

Using both purely design-based and model-assisted arguments, it is shown that, under conditions of high entropy, the variance of the Horvitz-Thompson (HT) estimator depends almost entirely on first order inclusion probabilities. Approximate expressions and estimators are derived for this "high entropy" variance of the HT estimator. Monte Carlo simulation studies are conducted to examine the statistical properties of the proposed variance estimators.

Key Words: Horvitz-Thompson estimator; Model assisted survey sampling; Monte Carlo simulation; Variance estimation.

1. Introduction

Let U denote a finite population of N units labeled $i = 1, ..., N$, and let Y_i denote the value for the i^{th} unit of a certain characteristic y. Consider the problem of estimating the population total $Y_\bullet = \sum_{i=1}^{N} Y_i$. If a sample, s, of n units is drawn without replacement from U with first order inclusion probabilities π_i, $i \in U$, the Horvitz-Thompson (HT) (1952) estimator of the total is $\hat{Y}_{\bullet\text{HT}} = \sum_{i \in s} Y_i \pi_i^{-1}$. In this paper, we confine consideration to fixed size sampling designs. For this important special case, Sen (1953) and Yates and Grundy (1953) showed independently that $\hat{Y}_{\bullet\text{HT}}$ has the variance

$$\mathrm{V}(\hat{Y}_{\bullet\text{HT}}) = (1/2) \sum_{i \in U} \sum_{j(\neq i) \in U} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2, \quad (1)$$

where π_{ij} is the second order or joint inclusion probability of the i^{th} and j^{th} population unites together in the same sample. They therefore suggested the variance estimator

$$\hat{\mathrm{V}}_{\text{SYG}}(\hat{Y}_{\bullet\text{HT}}) = (1/2) \sum_{i \in s} \sum_{j(\neq i) \in s} \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij}) (Y_i \pi_i^{-1} - Y_j \pi_j^{-1})^2. \quad (2)$$

This is known to perform better than the variance estimator proposed by Horvitz and Thompson (1952) (the latter, however, usually being unbiased for random n), but the critical dependence of (2) on π_{ij} has proved problematical (Brewer 1999). If one or more of the $N(N-1)/2$ distinct values of π_{ij} are zero, the estimator (2) is biased downwards. And if any of them should be very small compared with their corresponding values of $\pi_i \pi_j$, (2) will be unstable (that is, it will itself be subject to high variance). In addition, the double sum feature of (2) is quite inconvenient, especially for large sample sizes. Not only are there many more π_{ij}'s that there are π_i's; it is also frequently the case that the individual π_{ij}'s are problematic to evaluate. In view of these difficulties, the aim of this paper is to provide alternative variance estimators, which do not depend on the π_{ij}'s and are simple to compute.

In the next section, a new expression for the design-variance of the HT estimator is presented. This new expression leads, under high entropy conditions, to the derivation of an approximate formula for $\mathrm{V}(\hat{Y}_{\bullet\text{HT}})$, which is π_{ij}–free. In section 3, we check the usefulness of our approximate formulae using a model assisted approach. An estimator of our approximate variance is proposed in section 4; this variance estimator is expected to perform well under conditions of high entropy (meaning the absence of any detectable pattern or ordering in the selected sample units). Most sample selection schemes though, result in the selection of high entropy samples. With the aim of testing the usefulness of the variance estimator presented in section 4, some empirical studies were conducted. The main findings from these studies are reported in section 5. Some concluding remarks are provided in section 6.

2. Some Approximate Formulae for the Design-Variance of the HT Estimator

We begin this section by presenting an alternative formulation for the variance of the HT estimator, valid only when the sampling design is of fixed size. Before proceeding, we state the following relations, which will be useful later:

$$\sum_{j(\neq i) \in U} \pi_{ij} = (n-1)\pi_i, \quad i \in U \quad (3)$$

$$\sum_{j(\neq i) \in U} \pi_i \pi_j = (n - \pi_i)\pi_i, \quad i \in U \quad (4)$$

$$\sum_{i \in U} \sum_{j(\neq i) \in U} \pi_{ij} = n(n-1) \quad (5)$$

1. Ken Brewer, School of Finance and Applied Statistics, Australian National University, ACT 0200, Australia. E-mail: Ken.Brewer@anu.edu.au; Martin E. Donadio, Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia. E-mail: M.Donadio@abs.gov.au.

$$\sum_{i\in U}\sum_{j(\neq i)\in U}\pi_i\pi_j = n^2 - \sum_{i\in U}\pi_i^2 \qquad (6)$$

The alternative formulation is obtained as follows. We start with a trivial modification of (1),

$$\begin{aligned}V(\hat{Y}_{\bullet\mathrm{HT}}) &= (1/2)\sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij}) \\ &\quad \left\{\begin{array}{l}(Y_i\pi_i^{-1} - Y_\bullet n^{-1}) \\ -(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\end{array}\right\}^2 \\ &= (1/2)\sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij}) \\ &\quad \left\{\begin{array}{l}(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 + (Y_j\pi_j^{-1} - Y_\bullet n^{-1})^2 \\ -2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet n^{-1})\end{array}\right\}.\end{aligned}$$

Using the relations (3) and (4), the above equation may be shown to be identical to

$$\begin{aligned}V(\hat{Y}_{\bullet\mathrm{HT}}) &= \sum_{i\in U}\pi_i(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 \\ &\quad - \sum_{i\in U}\pi_i^2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 \\ &\quad - \sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij}) \\ &\quad (Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet^{-1}). \qquad (7)\end{aligned}$$

The first term in (7) is virtually the same as the variance of the corresponding Hansen-Hurwitz (1943) estimator of total for sampling at n draws with replacement, the probability of selecting unit i at each draw being $p_i = \pi_i/n, i\in U.$ The second term can be viewed as a finite population correction. Consequently, these two terms together plausibly constitute a first approximation to the entire variance of the HT estimator and, importantly, neither of them depends on the π_{ij}'s.

The magnitude of the third term depends mostly on the sampling design $p(s)$. Thus, if $p(s)$ is such that $\pi_{ij} \approx \pi_i\pi_j,$ for all $i\neq j\in U,$ then we can expect a very small third term in (7) (compared with the other two). This condition seems to be satisfied by high entropy sampling designs. For example, in simple random sampling without replacement (*srswor*), which maximizes the entropy among all fixed sized designs (see Hájek 1981), the second order inclusion probabilities can be written as $\pi_{ij} = \pi_i\pi_j[N(n-1)/\{n(N-1)\}]$. The factor $N(n-1)/\{n(N-1)\}$ is less than 1, and tends to 1 for large population and sample sizes. For this design, the third term in (7) accounts for only $1/N$ of the entire variance of the HT estimator. Furthermore, for several probability proportional-to-size designs, such as rejective sampling (Hájek 1964) and randomized systematic πps sampling (Hartley and Rao 1962), the condition $\pi_{ij} \approx \pi_i\pi_j$ also holds, provided N and n are large enough.

There are some exceptions, however, in which the third term in (7) can be important. The most important of these exceptions is systematic sampling from a population in which the units are arranged in a meaningful order prior to the selection. In such a case, a number of second order inclusion probabilities can even be equal to zero. This and other special cases need to be dealt with separately, and are not discussed further in this paper.

The rest of this section is devoted to deriving an approximation to $V(\hat{Y}_{\bullet\mathrm{HT}})$ that uses first order inclusion probabilities only. We start by proposing a simple approximation to the π_{ij} of the form

$$\pi_{ij} \cong \tilde{\pi}_{ij} = \pi_i\pi_j(c_i + c_j)/2, \quad i\neq j\in U. \qquad (8)$$

Three possible choices for $c_i, i\in U,$ are then:

$$c_i = (n-1)/(n-\pi_i), \qquad (9)$$

$$c_i = c = (n-1)\Big/\left(n - n^{-1}\sum_{k\in U}\pi_k^2\right) \text{ and} \qquad (10)$$

$$c_i = (n-1)\Big/\left(n - 2\pi_i + n^{-1}\sum_{k\in U}\pi_k^2\right). \qquad (11)$$

The first two choices of c_i are prompted by ratios of sums of π_{ij} to the corresponding sums of $\pi_i\pi_j$. Thus, on the one hand, formula (9) is obtained by comparing (3) with (4). On the other hand, formula (10) is suggested by the comparison of (5) and (6). Finally, formula (11) is based on the asymptotic expressions for π_{ij} obtained by Hartley and Rao (1962) and by Asok and Sukhatme (1976) for randomized systematic πps sampling and for Sampford's (1967) procedure respectively. To order $O(n^3N^{-3})$, both these asymptotic expressions simplify to

$$\begin{aligned}\tilde{\pi}_{ij} &= \pi_i\pi_j\{(n-1)/n\} \\ &\quad \left\{1 + n^{-1}(\pi_i + \pi_j) - n^{-2}\sum_{k\in U}\pi_k^2\right\},\end{aligned}$$

which in turn implies $c_i = \{(n-1)/n\}(1 + 2n^{-1}\pi_i - n^{-2}\sum_{k\in U}\pi_k^2)$. Under *srswor*, however, this choice of c_i does not yield the exact formula for the π_{ij}'s. For this reason, the slightly different expression given by (11) is used here, $(1 - 2n^{-1}\pi_i + n^{-2}\sum_{k\in U}\pi_k^2)$ being the first two terms in the Taylor expansion of the reciprocal of $(1 + 2n^{-1}\pi_i - n^{-2}\sum_{k\in U}\pi_k^2)$ and *vice versa*.

The next step consists of replacing the π_{ij}'s in the third term of (7) by the approximation (8). This replacement yields

$$\begin{aligned}&- \sum_{i\in U}\sum_{j(\neq i)\in U}(\pi_i\pi_j - \pi_{ij})(Y_i\pi_i^{-1} - Y_\bullet n^{-1}) \\ &\quad (Y_j\pi_j^{-1} - Y_\bullet n^{-1}) \\ &\cong - \sum_{i\in U}\sum_{j(\neq i)\in U}\pi_i\pi_j[1 - (c_i + c_j)/2] \\ &\quad (Y_i\pi_i^{-1} - Y_\bullet n^{-1})(Y_j\pi_j^{-1} - Y_\bullet n^{-1}) \\ &= \sum_{i\in U}(1 - c_i)\pi_i^2(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2,\end{aligned}$$

and thus the variance of the HT estimator may be approximated by

$$\begin{aligned}\tilde{\mathrm{V}}(\hat{Y}_{\bullet\mathrm{HT}}) &= \sum_{i\in U} [\pi_i - \pi_i^2 + (1-c_i)\pi_i^2](Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 \\ &= \sum_{i\in U} \pi_i(1-c_i\pi_i)(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2. \end{aligned} \quad (12)$$

This approximate variance has a very simple form. It is also without error under *srswor* for all the three choices of c_i presented above.

3. A Model Assisted Check on the Usefulness of the Approximate Variance Formulae

Consider the following ratio model as a possible description of the population being sampled:

$$\begin{aligned}&\xi: Y_i = \beta\pi_i + \varepsilon_i;\ \mathrm{E}_\xi\varepsilon_i = 0;\ \mathrm{E}_\xi\varepsilon_i^2 = \sigma_i^2; \\ &E_\xi(\varepsilon_i\varepsilon_j) = 0,\ i\neq j;\ i,\ j\in U. \end{aligned} \quad (13)$$

This is a shorthand model. It is intended to reflect the situation where the expected values of the Y_i are *intrinsically* proportional to the values X_i of an auxiliary variable x, and the inclusion probabilities π_i are *chosen* to be proportional to the X_i. It is of course impossible for the Y_i to be directly dependent on the inclusion probabilities as such, since those probabilities may be set quite arbitrarily by the person designing the sample.

The prediction or model expectation under ξ of the approximate variance expression (12) is

$$\begin{aligned}&\mathrm{E}_\xi\tilde{\mathrm{V}}(Y_{\bullet\mathrm{HT}}) \\ &= \mathrm{E}_\xi\sum_{i\in U}\pi_i(1-c_i\pi_i)(Y_i\pi_i^{-1} - Y_\bullet n^{-1})^2 \\ &= \mathrm{E}_\xi\sum_{i\in U}\pi_i(1-c_i\pi_i)(\varepsilon_i\pi_i^{-1} - \varepsilon_\bullet n^{-1})^2 \\ &= \sum_{i\in U}\sigma_i^2\left\{\begin{array}{r}\pi_i^{-1} - n^{-1} - c_i(1-2n^{-1}\pi_i) \\ -n^{-2}\sum_{k\in U}c_k\pi_k^2\end{array}\right\}, \end{aligned} \quad (14)$$

where $\varepsilon_\bullet = \sum_{i\in U}\varepsilon_i$. Ideally, expression (14) should be equal to $\mathrm{E}_\xi\,\mathrm{V}(\hat{Y}_{\bullet\mathrm{HT}})$, namely $\sum_{i\in U}\sigma_i^2(\pi_i^{-1}-1)$ (Godambe 1955; Godambe and Joshi 1965). This condition leads to the implicit formula

$$c_i = \left(1 - n^{-1} - n^{-2}\sum_{k\in U}c_k\pi_k^2\right)\Big/(1-2n^{-1}\pi_i),$$

which can be solved for c_i iteratively, starting with the trial value $c_i^{[1]} = (n-1)/n$. To $O(N^{-1})$, this iterative solution is identical to (11). Alternatively, a closed expression can be derived by putting (14) equal to $\sum_{i\in U}\sigma_i^2(\pi_i^{-1}-1)$ and then requiring that $c_i = c$ for all $i\in U$, in which case we obtain

$$c = (1-n^{-1})\sum_{i\in U}\sigma_i^2\Big/\sum_{i\in U}\sigma_i^2\left(\begin{array}{r}1-2n^{-1}\pi_i \\ -n^{-2}\sum_{k\in U}\pi_k^2\end{array}\right). \quad (15)$$

Under *srswor*, (15) becomes $c = N(n-1)/\{n(N-1)\}$, which yields the exact expression for $\mathrm{V}(\hat{Y}_{\bullet\mathrm{HT}})$. Even without *srswor*, replacing σ_i^2 by $\sigma^2\pi_i$ in (15) returns (10) for c. It is reassuring that the purely design-based analysis and the model-assisted one produce results in such close agreement.

4. Estimating the Design-Variance of the HT Estimator

The aim of this section is to propose a plausible sample estimator for the approximate design-variance of the HT estimator given in (12). One such estimator is

$$\hat{\tilde{\mathrm{V}}}(\hat{Y}_{\bullet\mathrm{HT}}) = \sum_{i\in s}(c_i^{-1} - \pi_i)(Y_i\pi_i^{-1} - \hat{Y}_{\bullet\mathrm{HT}}n^{-1})^2, \quad (16)$$

which is arrived at by replacing each population sum in (12) by the corresponding HT estimator, and adjusting by the factor c_i^{-1}. This estimator has some attractive properties: (i) For all three choices of c_i, it reduces to the standard variance estimator in the case of *srswor*; (ii) it is simple to compute, since no double sums are involved; and (iii) using Taylor linearization technique, it can be shown that (16) is approximately design-unbiased for (12).

A further attractive property of the estimator (16) is the following. When c_i is specified by (9), we have

$$\begin{aligned}c_i^{-1} - \pi_i &= (n-\pi_i)/(n-1) - \pi_i \\ &= \{n/(n-1)\}(1-\pi_i). \end{aligned} \quad (17)$$

The factor $(1-\pi_i)$ is easily interpretable as a finite population correction, while the factor $n/(n-1)$ has an entirely different role, which can be explained as follows. It is easy to see that $\hat{\beta} = \hat{Y}_{\bullet\mathrm{HT}}n^{-1}$ is a model unbiased estimator of β in model (13). Let us write $\hat{\sigma}_i^2 = (Y_i - \hat{\beta}\pi_i)^2$, for all i. Then $(Y_i\pi_i^{-1} - \hat{Y}_{\bullet\mathrm{HT}}n^{-1})^2 = (Y_i - \hat{\beta}\pi_i)^2\pi_i^{-2} = \hat{\sigma}_i^2\pi_i^{-2}$, $i\in U$. It is not difficult to show that the factor $n/(n-1)$ removes the (model) bias from $\sum_{i\in s}(Y_i\pi_i^{-1} - \hat{Y}_{\bullet\mathrm{HT}}n^{-1})^2 = \sum_{i\in s}\hat{\sigma}_i^2\pi_i^{-2}$ as an estimator of $\sum_{i\in s}\sigma_i^2\pi_i^{-2}$.

The choice of (9) to specify the value of c_i also renders particularly simple the calculation both of the HT estimate itself and of its estimated variance; for substituting (17) into (16) and expanding that expression into individual terms we obtain:

$$\hat{\tilde{\mathrm{V}}}(\hat{Y}_{\bullet\mathrm{HT}}) = \{n/(n-1)\}\left\{\begin{array}{l}\sum_{i\in s}Y_i^2\pi_i^{-2} - n^{-1}\hat{Y}_{\bullet\mathrm{HT}}^2 - \sum_{i\in s}Y_i^2\pi_i^{-1} \\ +2n^{-1}\hat{Y}_{\bullet\mathrm{HT}}\sum_{i\in s}Y_i - n^{-2}\hat{Y}_{\bullet\mathrm{HT}}^2\sum_{i\in s}\pi_i\end{array}\right\}.$$

This formula involves six expressions, namely $n, \hat{Y}_{\bullet\mathrm{HT}}$, $\sum_{i\in s} Y_i^2 \pi_i^{-2}$, $\sum_{i\in s} Y_i^2 \pi_i^{-1}$, $\sum_{i\in s} Y_i$, and $\sum_{i\in s} \pi_i$, which are the sample sums of 1 (unity), $Y_i \pi_i^{-1}$, $Y_i^2 \pi_i^{-2}$, $Y_i^2 \pi_i^{-1}$, Y_i, and π_i respectively. If these individual terms are cumulated over every sample unit, then $\hat{Y}_{\bullet\mathrm{HT}}$ and $\hat{\mathrm{V}}(\hat{Y}_{\bullet\mathrm{HT}})$ can be evaluated together, using only a single pass of the sample data.

Note that, if non-response is present, a first order correction for it may be obtained by conditioning the sample on the achieved sample size, which we may denote here by n'. That would involve replacing the original first order inclusion probabilities, π_i, by the "adjusted inclusion probabilities", $\pi_i' = \pi_i n'/n$. (This terminology has been taken from Furnival, Gregoire and Grosenbaugh (1987), where the same type of adjustment was used in a different context). The summations over the achieved sample s', would then be n', $\sum_{i\in s'} Y_i \pi_i'^{-1}$, $\sum_{i\in s'} Y_i^2 \pi_i'^{-2}$, $\sum_{i\in s'} Y_i^2 \pi_i'^{-1}$, $\sum_{i\in s'} Y_i$, and $\sum_{i\in s'} \pi_i'$ respectively.

Beyond the properties listed above, a further study of (16) is possible with the aid of the model ξ of (13). The most desirable expression for the ξ–expectation of an estimator of $\mathrm{V}(\hat{Y}_{\bullet\mathrm{HT}})$ is $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$, because this in turn has design-expectation $\sum_{i\in U} \sigma_i^2 (\pi_i^{-1} - 1)$, which is the lower bound for the anticipated variance of any unbiased estimator (Godambe 1955; Godambe and Joshi 1965). For all the three definitions of c_i, the $\xi-$expectation of (16) differs from $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$ by terms of order $O(Nn^{-1})$. Although these "unwanted" terms have opposite signs and therefore tend to cancel, they are not entirely negligible, being only $O(N^{-1})$ smaller than the variance itself.

In view of this, a new version of c_i, which retained the (design) properties (i)–(iii) for (16) and provided a closer expression to $\sum_{i\in s} \sigma_i^2 \pi_i^{-1}(\pi_i^{-1} - 1)$ for the ξ–expectation of (16), was desirable. These requirements are satisfied by a c_i defined as follows:

$$c_i = \frac{(n-1)}{\left\{ n - (2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1} \sum_{k\in U} \pi_k^2 \right\}}, \quad (18)$$

for all $i \in U$. With this definition of c_i, the ξ–expectation of (16) still contains some "unwanted" terms, but they now consist only of a single term of order $O(Nn^{-2})$–which is therefore smaller than $\mathrm{V}(\hat{Y}_{\bullet\mathrm{HT}})$ by a factor of order $O(N^{-1}n^{-1})$–and other terms of smaller magnitude still.

5. Some Empirical Results

With the aim of evaluating the performance of the variance estimator proposed in section 4, some empirical studies were conducted. Three other variance estimators were also included in these studies: (i) the SYG estimator, given in (2); (ii) the variance estimator suggested by Hájek (1964, page 1520),

$$\hat{\mathrm{V}}_{\mathrm{HAJ}}(\hat{Y}_{\bullet\mathrm{HT}}) = \{n/(n-1)\} \sum_{i\in s} (1-\pi_i)(Y_i \pi_i^{-1} - A_s)^2, \quad (19)$$

where $A_s = \sum_{i\in s} a_i Y_i \pi_i^{-1}$, $a_i = (1-\pi_i)/\sum_{k\in s}(1-\pi_k)$; and (iii) a slight modification of (19) proposed by Deville (1999),

$$\hat{\mathrm{V}}_{\mathrm{DEV}}(\hat{Y}_{\bullet\mathrm{HT}}) = \frac{1}{1-\sum_{i\in s} a_i^2} \sum_{i\in s} (1-\pi_i)(Y_i \pi_i^{-1} - A_s)^2. \quad (20)$$

It is worth mentioning that the estimator (19) was originally intended only for a particular high entropy design, namely rejective sampling, and not for all the high entropy ones. Later on, however, this estimator was proposed for its use with some other high entropy designs. For example, Rosén (1997) suggested the use of (19) in the context of Pareto sampling.

The inclusion of the estimators (2), (19) and (20) in our impirical studies deserves a brief explanation. The SYG variance estimator would usually be the preferred choice if the π_{ij} were known and were neither zero nor very small compared with the corresponding $\pi_i \pi_j$. Under these conditions, it would then be natural to ask: Is there a significant difference, in terms of performance, between (2) and the simpler estimator (16)? On the other hand, a comparison with (19) and (20) is of interest because these two estimators share with (16) the simplicity and π_{ij} – free features. Thus, they are "competitors" in the same class.

The performance of a variance estimator can be assessed in different ways; here we will focus on *bias* and *stability*. The main findings from our studies are reported in the remainder of this section. We will consider two cases separately, namely $n = 2$ and $n > 2$.

5.1 Case $n = 2$

With the aim of testing the variance estimators under different situations, nine small populations were used in this study, most of which were also included in the stability studies carried out by Rao and Bayless (1969). Table 1 summarizes the main features of each population, including the coefficients of variance (CV) of y and x, and the correlation coefficient, ρ, between y and x. Here, y is the variable for which total estimates are sought, and x is an auxiliary variable that may be used for sample selection. Note that N varies from 10 to 20, $\mathrm{CV}(x)$ from 0.14 to 0.73, and ρ from 0.19 to 0.94. This provides a good mixture of populations with different characteristics.

The inclusion probabilities are chosen to proportional to x, *i.e.*, $\pi_i = 2X_i / X_\bullet$, for all i. Two sampling designs are considered here, namely Brewer's (1963) procedure (BREWER) and Tillé's (1996) elimination procedure (TILLÉ). For both procedures, the π_{ij} are simple to compute and, for these nine populations, they are strictly positive (this condition is not always satisfied by TILLÉ). Moreover, since $n = 2$, for any sample $s = \{i, j\}$ we have $p(s) = \pi_{ij}$. Hence we can obtain the exact statistical properties of any given variance estimator $\hat{\mathrm{V}}$.

To this end, let S denote the set of all possible samples of size $n = 2$ from a population U. The expectation of $\hat{\mathrm{V}}$ is then defined as

$$E(\hat{V}) = \sum_{s \in S} p(s)\hat{V}(s),$$

and its standard error (SE) as

$$\mathrm{SE}(\hat{V}) = \left\{ \sum_{s \in S} p(s)[\hat{V}(s) - E(\hat{V})]^2 \right\}^{1/2}.$$

For each of the two sampling designs mentioned above, Table 2 displays the *relative bias* $\mathrm{RB}(\hat{V}) = \mathrm{E}(\hat{V}) / \mathrm{V}(\hat{Y}_{\bullet \mathrm{HT}}) - 1$, expressed as a percentage, of the six π_{ij} – free variance estimators. The first two of these estimators need no explanation; the other four correspond to (16) coupled with (9), (10), (11), and (18) respectively. Since for $n = 2$ (only), $\hat{V}_{\mathrm{DEV}}$ and $\hat{V}_{16.9}$ are identical, they both appear in the same row. In order to simplify the reading of the table, the smallest RB (in absolute terms) in each population and sampling design has been highlighted.

The results in Table 2 prompt the following comments: (i) the performance of the π_{ij} –free variance estimators is reasonably good for all populations, with the possible exception of Population 4. An examination of the relationship between *x* and *y* for this population reveals the presence of some curvature, with larger cities growing at a higher rate. There is also an outlier–city 26–for which the number of people almost tripled in the 10–year period between 1920 and 1930. Another interesting case is given by Populations 5 and 6. These two populations have identical definitions, thus one would expect to obtain similar results for them. However, the RB figures for Population 5 are considerably worse than those for Population 6, specially for BREWER. The only noticeable difference between these two populations is that Population 5 contains an outlier (Farm 14 in the reference provided). It would appear then that the presence of outliers may result in some additional bias in these variance estimators. (ii) The estimator $\hat{V}_{16.18}$ seems to be the best of the class, performing remarkably well in all situations, and showing the smallest bias figures (in absolute values) in most cases; (iii) The estimator $\hat{V}_{16.10}$ tends to exhibit the largest bias figures.

Regarding stability, Table 3 reports the *coefficient of variation* $\mathrm{CV}(\hat{V}) = \mathrm{SE}(\hat{V}) / \mathrm{E}(\hat{V})$, expressed as a percentage, of all the seven variance estimators. It can be seen that the π_{ij} –free variance estimators tend to be more efficient (lower CVs) than $\hat{V}_{\mathrm{SYG}}$, although the gains are small. Otherwise, there is little to choose from among these variance estimators, even though $\hat{V}_{16.10}$ is the best performer in all but the last population.

Table 1
Description of the Nine Small Populations

Pop.	*Source*	*y*	*x*	*N*	CV(*y*)	CV(*x*)	ρ
1	Cochran (1963, page 325)	No. of persons per block	No. of rooms per block	10	0.15	0.14	0.65
2	Yates (1981, page 150) Kraals 26–38	No. of persons absent	Total no. of persons	13	0.67	0.47	0.72
3	Rao (1963, page 207)	Corn acreage in 1960	Corn acreage in 1958	14	0.39	0.43	0.93
4	Cochran (1963, page 156) Cities 19–33	No. of people in 1930	No. of people in 1920	15	0.67	0.69	0.94
5	Sampford (1962, page 61) Even units	Oat acreage in 1957	Total acreage in 1947	17	0.61	0.71	0.80
6	Sampford (1962, page 61) Odd units	Oat acreage in 1957	Total acreage in 1947	18	0.75	0.73	0.91
7	Yates (1981, page 153)	Vol. of timber	Eye-estimated vol. of timber	20	0.51	0.48	0.49
8	Sukhatme (1954, page 279) Circles 1–20	Wheat acreage	No. of villages	20	0.63	0.50	0.59
9	Horvitz and Thompson (1952, page 682)	No. of households	Eye-estimated no. of households	20	0.44	0.40	0.87

Table 2
RB (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	$\hat{V}_{\mathrm{HAJ}}$	−1.04	−2.97	−2.60	−6.05	−3.64	0.08	−0.81	−1.48	1.13
R	$\hat{V}_{\mathrm{DEV}}$, $\hat{V}_{16.9}$	−0.98	−2.52	−2.29	−5.21	−3.00	0.54	−0.63	−1.33	1.24
E	$\hat{V}_{16.10}$	−1.37	−3.55	−3.21	−7.16	−4.31	0.82	−0.94	−1.89	1.80
W	$\hat{V}_{16.11}$	−0.59	−1.49	−1.37	−3.26	−1.69	0.26	−0.31	−0.76	0.68
.	$\hat{V}_{16.18}$	**−0.20**	**−0.46**	**−0.46**	**−1.31**	**−0.38**	**−0.01**	**0.00**	**−0.19**	**0.13**
T	$\hat{V}_{\mathrm{HAJ}}$	−1.06	−4.40	−1.07	−5.90	−1.86	−0.41	0.32	−1.10	0.82
I	$\hat{V}_{\mathrm{DEV}}$, $\hat{V}_{16.9}$	−1.00	−3.94	−0.75	−5.03	−1.19	**0.07**	0.51	−0.95	0.93
L	$\hat{V}_{16.10}$	−1.39	−4.91	−1.68	−6.91	−2.47	0.33	**0.19**	−1.50	1.48
L	$\hat{V}_{16.11}$	−0.62	−2.98	**0.17**	−3.14	**0.09**	−0.20	0.83	−0.39	0.38
É	$\hat{V}_{16.18}$	**−0.23**	**−2.02**	1.10	**−1.25**	1.37	−0.46	1.15	**0.17**	**−0.17**

Table 3
CV (%) of Variance Estimators for $n = 2$

		Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
B	$\hat{V}_{SYG}$	123	126	118	245	138	127	158	127	**133**
R	$\hat{V}_{HAJ}$	121	119	115	238	131	125	155	124	134
E	$\hat{V}_{DEV}$, $\hat{V}_{16.9}$	121	119	115	238	131	125	155	124	134
W	$\hat{V}_{16.10}$	**120**	**117**	**114**	**236**	**128**	**124**	**153**	**123**	135
E	$\hat{V}_{16.11}$	122	122	116	241	133	126	157	125	133
R	$\hat{V}_{16.18}$	122	125	117	243	136	127	158	126	133
T	$\hat{V}_{SYG}$	123	143	118	248	147	131	164	131	134
I	$\hat{V}_{HAJ}$	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{DEV}$, $\hat{V}_{16.9}$	121	118	115	238	128	125	155	124	134
L	$\hat{V}_{16.10}$	**121**	**116**	**114**	**235**	**125**	**124**	**154**	**123**	135
É	$\hat{V}_{16.11}$	122	121	115	240	130	125	157	125	133
	$\hat{V}_{16.18}$	122	123	116	243	133	126	159	126	**133**

5.2 Case $n > 2$

In this section, we adopt a standard Monte Carlo simulation approach to examine the performance of the variance estimators. Two real populations are used in this study. The first one is a population of 220 blocks (BL220) taken from Appendix E in Kish (1965). The dataset contains two variables: Y_i = no. of dwellings occupied by renters in block i, and X_i = total no. of dwellings in block i. Some features of this population are: $\mathrm{CV}(y) = 1.05$, $\mathrm{CV}(x) = 0.85$, and $\rho = 0.97$.

The second population comprises 281 municipalities (MU281), and is given in Särndal, Swensson, and Wretman (1992). The role of the study variable, y, is played by RMT85, revenues from the 1985 municipal taxation, while P75, the municipality population in 1975, is used as a measure of size. The main characteristics of this population are: $\mathrm{CV}(y) = 1.06$, $\mathrm{CV}(x) = 0.96$, and $\rho = 0.99$.

Samples of sizes $n = 10$, 20 and 40 with $\pi_i \propto X_i$, $i \in U$, are drawn from BL220 and MU281 by means of randomized systematic πps sampling (RANSYS) and TILLÉ. For each sample, we compute a total estimate using the HT estimator, and variance estimates using the seven variance estimators mentioned in the previous section (for RANSYS, however, the Hartley and Rao (1962) approximation to the π_{ij}, is used in formula (2)). This sampling-estimation process is repeated $R = 50{,}000$ times.

Table 4 shows the observed Monte Carlo relative biases of the variance estimators for RANSYS and TILLÉ. Note that, for TILLÉ, no values have been provided in the row corresponding to the SYG variance estimator. This is because, given the populations, measures of size, and sample sizes employed here, TILLÉ produces strictly positive π_{ij}, which means that the SYG variance estimator is design unbiased. All the figures in this table are reasonably small, which seems to support our belief that, under conditions of high entropy, the calculation of the π_{ij} is not essential for obtaining nearly unbiased variance estimates. Within the group of π_{ij}–free estimators, there are no noticeable differences among them so far as RANSYS is concerned, but $\hat{V}_{HAJ}$ and its relative, $\hat{V}_{DEV}$, seem to perform somewhat better than the $\hat{V}_{16.*}$ family so far as TILLÉ is concerned, especially for $n = 40$. However, all the observed TILLÉ biases are positive and tend to increase as the sample size increases. This seems to indicate that TILLÉ is slightly lower in entropy than RANSYS, in which case the higher observed biases for the $\hat{V}_{16.*}$ family are reflecting the actual facts quite accurately.

Table 4
RB (%) of Variance Estimators for $n > 2$

Variance estimators	RANSYS			TILLÉ		
	$n = 10$	$n = 20$	$n = 40$	$n = 10$	$n = 20$	$n = 40$
			BL220			
$\hat{V}_{SYG}$	0.13	1.02	**– 0.27**	–	–	–
$\hat{V}_{HAJ}$	– 0.14	**0.47**	– 2.35	1.49	**2.18**	**3.27**
$\hat{V}_{DEV}$	– 0.12	0.54	– 2.15	1.52	2.25	3.48
$\hat{V}_{16.9}$	**– 0.06**	0.83	– 0.52	1.58	2.54	5.21
$\hat{V}_{16.10}$	– 0.23	0.64	– 0.75	**1.41**	2.34	4.97
$\hat{V}_{16.11}$	0.11	1.02	– 0.30	1.75	2.73	5.45
$\hat{V}_{16.18}$	0.13	1.03	– 0.29	1.77	2.74	5.45
			MU281			
$\hat{V}_{SYG}$	– 0.27	– 0.43	0.77	–	–	–
$\hat{V}_{HAJ}$	– 0.40	– 0.75	– 0.59	0.64	**1.01**	**1.93**
$\hat{V}_{DEV}$	– 0.37	– 0.68	**– 0.39**	0.67	1.09	2.14
$\hat{V}_{16.9}$	– 0.34	– 0.51	0.67	0.70	1.26	3.22
$\hat{V}_{16.10}$	– 0.40	– 0.58	0.58	**0.63**	1.19	3.13
$\hat{V}_{16.11}$	**– 0.27**	– 0.43	0.76	0.77	1.34	3.31
$\hat{V}_{16.18}$	– 0.27	**– 0.43**	0.76	0.78	1.34	3.32

In order to test whether TILLÉ is of slightly lower entropy than RANSYS or not, we compared their Monte

Carlo variances (MCV) with formula (12), the high entropy approximation to the HT variance. The most accurate version of c_i, that is (18), was used to compute (12). The comparison is presented in Table 5. It is seen that the TILLÉ variances are somewhat smaller than the corresponding RANSYS variances. Moreover, the approximate variances provided by (12) are in closer agreement with the RANSYS variances. These findings support our previous conjecture that the entropy for TILLÉ is slightly lower than that for RANSYS, particularly when the finite population correction is appreciable.

Table 5
Comparison of Variances (all values in 10^4)

	BL220			MU281		
	$n=10$	$n=20$	$n=40$	$n=10$	$n=20$	$n=40$
(12) + (18)	14.06	6.572	2.830	565.5	264.3	113.7
MCV – RANSYS	14.07	6.520	2.841	566.2	265.3	112.8
MCV – TILLÉ	13.87	6.404	2.691	560.0	257.6	108.9

Next we focus on stability. Table 6 reports the observed Monte Carlo SE of the variance estimators. Clearly, there are no differences worth mentioning among the variance estimators. The same is true for a comparison of the two sampling procedures. It seems that stability does not constitute a relevant factor when choosing between these variance estimators.

Table 6
CV (%) of Variance Estimators for $n=2$

Variance estimators	RANSYS			TILLÉ		
	$n=10$	$n=20$	$n=40$	$n=10$	$n=20$	$n=40$
			BL220			
$\hat{V}_{SYG}$	58.31	41.16	30.70	57.43	40.41	29.54
$\hat{V}_{HAJ}$	57.90	40.49	**29.48**	57.39	40.24	**29.08**
$\hat{V}_{DEV}$	57.90	40.49	29.48	57.39	40.24	29.08
$\hat{V}_{16.9}$	57.02	40.54	29.64	57.41	40.29	29.24
$\hat{V}_{16.10}$	**57.79**	**40.45**	29.56	**57.29**	**40.19**	29.16
$\hat{V}_{16.11}$	58.04	40.64	29.73	57.53	40.39	29.32
$\hat{V}_{16.18}$	58.05	40.65	29.73	57.55	40.39	29.32
			MU281			
$\hat{V}_{SYG}$	54.90	37.29	25.33	55.07	37.50	25.45
$\hat{V}_{HAJ}$	54.69	36.98	24.96	54.79	37.07	24.78
$\hat{V}_{DEV}$	54.68	36.98	24.95	54.79	37.07	24.77
$\hat{V}_{16.9}$	54.67	36.92	24.70	54.77	37.01	24.52
$\hat{V}_{16.10}$	**54.63**	**36.89**	**24.66**	**54.74**	**36.98**	**24.48**
$\hat{V}_{16.11}$	54.70	36.95	24.74	54.81	37.04	24.56
$\hat{V}_{16.18}$	54.71	36.96	24.74	54.81	37.04	24.56

6. Summary

Estimators are derived for what, in the context of any high entropy selection procedure, is a close approximation to the design variance of the HT estimator of a total.

These estimators resemble, but are not identical to other variance estimators suggested for certain particular high entropy selection procedures by Hájek (1964), Rosen (1997), and Deville (1999). All these estimators have the important advantage over the standard SYG variance estimator that their formulae do not involve the second order inclusion probabilities, π_{ij}.

Empirical investigations indicate that these estimators all behave acceptably well, both for the important special case $n=2$ and when n takes large values. The estimator given by (16) with c_i defined by (18), which has certain near-optimal theoretical properties, appears to be noticeably less biased than the others for $n=2$, but not for larger values of n.

For the case $n>2$, two high entropy procedures were used, namely systematic sampling from a randomly ordered population (RANSYS) and the procedure proposed by Tillé (1996) (TILLÉ). The biases in all the variance estimators TILLÉ than for RANSYS, and particularly so when n took its largest value of 40. The differences between the TILLÉ biases and the RANSYS biases were also positive for all values of n, and again particularly so when $n=40$. We conjecture that these differences may indicate that TILLÉ is a slightly lower entropy (and typically lower variance selection procedure than RANSYS).

Acknowledgements

The authors wish to thanks Dr. P.S. Kott for suggesting equation (10) in a private communication, and an anonymous referee for three other suggestions that have added value to this paper.

References

Asok, C., and Sukhatme, B.V. (1976). On sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

Brewer, K.R.W. (1999). Cosmetic calibration for unequal probability samples. *Survey Methodology*, 25, 205-212.

Cochran, W.G. (1963). *Sampling Techniques*. 2nd Ed. New York: John Wiley & Sons, Inc.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Furnival, G.M., Gregoire, T.G. and Grosenbaugh, L.R. (1987). Adjusted inclusion probabilities with 3P sampling. *Forest Science*, 33, 617-631.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*, 17, 269-278.

Godambe, V.P., and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I, II, and III. *Annals of Mathematical Statistics*, 36, 1707-1742.

Hájek, J. (1964). Assymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Hájek, J. (1981). Sampling from a finite population. New York: Marcel Dekker.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Rao, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

Rao, J.N.K., and Bayless, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Sampford, M.R. (1962). *An Introduction of Sampling Theory*. Edinburgh and London: Oliver and Boyd Ltd.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agriculture Statistics*, 5, 119-127.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications*. Ames, Iowa State College Press.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys*. 4th Ed. London: Charles Griffin and Co.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.