Social Statistics Research Centre
University of Southampton

A Simple Variance Estimator for Unequal Probability Sampling Without Replacement

Yves G. Berger

Abstract

Survey sampling textbooks often refer to the Sen-Yates-Grundy variance estimator for use with without replacement unequal probability designs. This estimator is rarely implemented, because of the complexity of determining joint inclusion probabilities. In practice, the variance is usually estimated by simpler variance estimators such as the Hansen-Hurwitz with replacement variance estimator; which often leads to overestimation of the variance for large sampling fraction that are common in business surveys. We will consider an alternative estimator: the Hájek (1964) variance estimator that depends on the first-order inclusion probabilities only and is usually more accurate than the Hansen-Hurwitz estimator. We review this estimator and show its practical value. We propose a simple alternative expression; which is as simple as the Hansen-Hurwitz estimator. We also show how the Hájek estimator can be easily implemented with standard statistical packages.

SSRC Methodology Working Paper M03/09

A Simple Variance Estimator for Unequal Probability Sampling Without Replacement

Yves G. Berger
Department of Social Statistics,
University of Southampton,
Southampton, SO17 1BJ
United Kingdom
E-mail: ygb@soton.ac.uk

May 20, 2003

Abstract

Survey sampling textbooks often refer to the Sen-Yates-Grundy variance estimator for use with without replacement unequal probability designs. This estimator is rarely implemented, because of the complexity of determining joint inclusion probabilities. In practice, the variance is usually estimated by simpler variance estimators such as the Hansen-Hurwitz with replacement variance estimator; which often leads to overestimation of the variance for large sampling fraction that are common in business surveys. We will consider an alternative estimator: the Hájek (1964) variance estimator that depends on the first-order inclusion probabilities only and is usually more accurate than the Hansen-Hurwitz estimator. We review this estimator and show its practical value. We propose a simple alternative expression; which is as simple as the Hansen-Hurwitz estimator. We also show how the Hájek estimator can be easily implemented with standard statistical packages.

Key words: Design-based inference, Hansen-Hurwitz variance estimator, Inclusion probabilities, π-estimator, Sen-Yates-Grundy variance estimator.

1 Introduction

Unequal probability sampling was first suggested by Hansen and Hurwitz (1943) in the context of with-replacement sampling. Narain (1951), Horvitz and Thompson (1952) developed the corresponding theory for sampling without replacement. Gabler (1984) shows the superiority of sampling without replacement over sampling with replacement. Variance estimation for sampling with-replacement is straightforward (Hansen & Hurwitz, 1943). However, for sampling without replacement, the design unbiased Sen-Yates-Grundy variance estimator (Sen,

1953; Yates and Grundy, 1953) is hard to compute because of joint inclusion probabilities. Although exact computation of these probabilities is possible with specific sampling designs like with the Chao (1982) sampling design, their calculation becomes practically impossible when the sample size is large. It is also inconceivable to provide these probabilities in released data-sets, as the set of joint inclusion probabilities is a series of $n(n-1)/2$ values; where n denotes the sample size. Moreover, standard statistical packages like SPSS®, SAS®, STATA® do not deal with these probabilities. Specialized software like SUDAAN® needs to be used. However, even SUDAAN® does not include actual computation of these probabilities. They need to be specified by the user.

The aim of this paper is to show that it is possible to estimate the sampling variance without computing joint inclusion probabilities by using the Hájek (1964) variance estimator. Our aim is to show the practical importance of this estimator and how it can be implemented using weighted least squares (WLS) regression, which is straightforward with standard statistical packages.

In Section 2, we review the issue of variance estimation. In Section 3, we introduce the Hájek variance estimator and we propose a simpler alternative expression. In Section 4, we introduce alternative variance estimator that are as simple as the Hájek variance estimator. In Section 5, the accuracy of the Hájek estimator is studied through Monte-Carlo studies.

2 Complexity of Variance Estimation

Consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ containing N units. Suppose we wish to estimate the population total

$$Y = \sum_{i \in U} y_i$$

where y_i is the value of a study variable of a unit labelled i. The *π-estimator* (Narain, 1951; Horvitz and Thompson, 1952) of Y is

$$\widehat{Y} = \sum_{i \in s} \breve{y}_i. \tag{1}$$

where s is a sample, $\breve{y}_i = y_i \pi_i^{-1}$ and where π_i is the first-order inclusion probabilities of unit i; that is, the probability for unit i to be sampled. The variance of the π-estimator plays an important role in variance estimation, as most estimators of interest can be linearized to involve π-estimators. The sampling variance of $\widehat{Y}$ for fixed sample size designs is given by

$$\sigma^2_Y = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j)\ \breve{y}_i \breve{y}_j. \tag{2}$$

π_{ij} is the joint inclusion probabilities of unit i and j; that is, the probability that both units i and j are selected.

A design unbiased estimator of σ_Y^2 is given by the *Sen-Yates-Grundy estimator*

$$\widehat{\sigma}_{YG}^2 = \frac{1}{2}\sum_{i\in s}\sum_{j\in s}(\pi_i\pi_j - \pi_{ij})\pi_{ij}^{-1}(\breve{y}_i - \breve{y}_j)^2. \tag{3}$$

The estimator (3) is hard to implement in practice, as the π_{ij} are often unknown except for special cases such as stratified simple random sampling (STSRS). Moreover, the double sum feature is computationally inconvenient for large samples. Furthermore, the computation of the π_{ij} requires the values of the π_i for all the units of the population, whereas it is common to known the value of π_i only for the sampled units. In this case, the π_{ij} cannot be computed. There are alternative methods (Smith, 2001)of variance estimation that do not involve π_{ij}, such as replication methods. In the next section, we show that the variance can be easily estimated without using computationally intensive methods like replication methods or any methods that would involve the actual computation of the π_{ij}.

3 The Hájek approach

The Hájek (1964) variance estimator can be interpreted as a modified Hansen-Hurwitz (1943) estimator (see (11) below) for sampling without replacement. The Hájek estimator is implemented by Statistics Sweden in the software `CLAN`® (Andersson and Nordberg, 1994) and by the French Office for Statistics (INSEE).

We suppose that the sampling design is a single stage stratified sampling design with unequal probabilities within each stratum. Let us denote the strata by $U_1, ..., U_H$. We suppose that a sample s_h of size n_h is selected without replacement within each stratum U_h of size N_h. In this paper, we use design-based arguments. For simplicity, we assume throughout that the sample data are free from errors due to non-response and from errors of measurement.

3.1 The Hájek Variance Approximation

Hájek (1981) proposed the following approximation:

$$\pi_{ij} \approx \pi_i\pi_j[1 - (1-\pi_i)(1-\pi_j)d_h^{-1}] \tag{4}$$

when $i \neq j \in U_h$ and where

$$d_h = \sum_{i\in U_h}\pi_i(1-\pi_i).$$

Hájek (1964, 1981) showed that this approximation is valid when rejective sampling is implemented in each stratum. Berger (1998) showed that this approximation can be used for a larger class of highly randomized or high entropy sampling designs; which includes the successive (Hájek, 1964) and the Rao-Sampford (Rao, 1965 and Sampford, 1967) sampling designs. The systematic

sampling design is not a high entropy sampling design. However, in Section 3.2, we show briefly how the Hájek variance estimator can be extended to accommodate this sampling design.

By substituting π_{ij} from (4) into (3), we obtain an estimator for the variance, which is however not suitable, as the approximation (4) and the double sum in (3) give unstable variance estimates. In order words, the Sen-Yates-Grundy estimator involving approximations of π_{ij} can be unstable. In Section 3.2, we propose an alternative method that consists of estimating an approximation to the variance.

By substituting (4) into (2), we obtain the following approximation to the variance

$$\sigma^2_{Haj} = \sum_{h=1}^{H} \sum_{i \in U_h} \pi_i (1 - \pi_i)(\breve{y}_i - B_h)^2$$

where

$$B_h = d_h^{-1} \sum_{i \in U_h} \pi_i (1 - \pi_i) \breve{y}_i.$$

An alternative expression for σ^2_{Haj} is

$$\sigma^2_{Haj} = \sum_{i \in U} c_i \; e_i^2 \tag{5}$$

where

$$\begin{aligned} e_i &= \breve{y}_i - B_h \qquad (i \in U_h); \\ c_i &= \pi_i (1 - \pi_i). \end{aligned}$$

An alternative expression for B_h is

$$B_h = \left(\sum_{j \in U} c_j \; z_{jh}^2 \right)^{-1} \sum_{i \in U} c_i \breve{y}_i \; z_{ih} \tag{6}$$

where $z_{ih} = 1$ if the $i \in U_h$ and otherwise $z_{ih} = 0$. The stratification variables z_{ih} are the indicator variables for the strata. It is useful to write B_h this way, as e_i can now be interpreted as WLS residuals of the *working regression*

$$\breve{y}_i = \sum_{h=1}^{H} B_h \; z_{ih} + e_i. \tag{7}$$

The term 'working regression' is used to emphasize that we do not assume that (7) is a super-population model. This regression is used to define e_i in (5). The fact that the e_i can be interpreted as residuals is a consequence of approximation (4) of the joint-inclusion probabilities. Obviously, the more the population is stratified, the more the population scatter conforms to a linear pattern (7), the smaller the population residuals e_i and the smaller the variance. If the population is not well described by (7), the improvement on the π-estimator may be modest, but (5) is still a good approximation to the variance.

3.2 The Hájek Variance Estimator

A natural estimator of σ^2_{Haj} is given by

$$\widehat{\sigma}^2_{Haj} = \sum_{i \in s} \breve{c}_i \ \widehat{e}_i^2 \tag{8}$$

where $\breve{c}_i = n_h(n_h - 1)^{-1}(1 - \pi_i)$ $(i \in U_h)$ is a factor including a finite population correction (FPC) and a correction for degrees of freedom (DF). The $\widehat{e}_i$ are the WLS residuals

$$\widehat{e}_i = \breve{y}_i - \sum_{h=1}^{H} \widehat{B}_h \ z_{ih} \tag{9}$$

where

$$\widehat{B}_h = \left(\sum_{j \in s} \breve{c}_j \ z_{jh}^2 \right)^{-1} \sum_{i \in s} \breve{c}_i \ \breve{y}_i \ z_{ih}.$$

Simple algebra establishes that (8) is algebraically equivalent to the stratum-by-stratum Hájek (1964, p. 1520). If we substitute π_i by n_h/N_h into (8) when $i \in U_h$, we obtain the usual stratum-by-stratum variance estimator for STSRS

$$\widehat{\sigma}^2_{Haj} = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{\widehat{S}_h^2}{n_h}$$

where $\widehat{S}_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_i - \overline{y}_h)$ and $\overline{y}_h = n_h^{-1} \sum_{i \in s_h} y_i$. Thus, when the units are selected with equal probabilities in each stratum, (8) equals (3). Thus, (8) is an approximation of (3) when we have varying first-order inclusion probabilities within strata.

In practice, (8) is simple to compute, as it does not require the π_{ij}. Moreover, (8) is stable as it is a simple sum that estimates a simple sum (5). If we know in which stratum each unit belongs, it is easy to specify the H stratification variables z_{ih}. As $\widehat{B}_h$ is the usual WLS estimate of a regression coefficient, any standard statistical packages can be used to compute the $\widehat{B}_h$ and the set of residuals $\widehat{e}_i$. The variance (8) is just a weighted sum of these residuals. The merit of this method is the fact that the variance estimator is only computed through a set of residuals and only requires the values of π_i for the sampled units.

The set of $(1 - \pi_i)$ in (8) can be viewed as generalised FPC. Indeed, with STSRS we have the usual FPC $1 - \pi_i = 1 - f_h$. A correction for DF $n_h(n_h - 1)^{-1}$ is also included in (8). There are other effects of the sampling design also included in (8). The effect of stratification is specified by the residuals $\widehat{e}_i$, as the working regression (7) uses the stratification variables as independent variables. This is the major effect of the sampling design. The effect of the π_i is also included in the residuals as the independent variables $\breve{y}_i$ in (7) is the study variable divided by the π_i. There are remainder effects not included in (8); which explains the slight differences in the variance (2) due to the method of sampling

used. The alternative expression (8) has the advantage of revealing the main effects of the sampling design due to: stratification, the unequal probabilities, the FPC and the correction for DF. This allows us to quantify the impact of these effects on the variance.

Although $\widehat{\sigma}^2_{Haj}$ is applicable under single stage stratified sampling designs, it can be generalised to more complex sampling designs. For example, with two stage designs, the variance is usually estimated (Skinner, 1989) by a variance between the primary sampling units (PSU). Thus, (8) can be implemented with i representing the PSU label and y_i an estimate of the total of the i-th PSU.

As already stated, (8) is suitable for high entropy sampling designs. Thus, (8) may not be suitable with systematic samples, as the entropy of this sampling design is low. Berger (2003a, 2003b) proposes an adjusted Hájek variance estimator for systematic sampling that includes additional independent variables in the working regression (7). For example, if U is composed of a single stratum, we have one additional independent variable given by

$$x_i = \pi_i^c \left(\sum_{q=1}^{n} I_{iq} - nI_{i1} \right) \tag{10}$$

where $\pi_i^c = \sum_{j \in U; j \leq i} \pi_j - \pi_i/2$ is the smooth cumulative sum of the π_i and I_{iq} is the indicator variables for the group $G_q = \{i \in u : q - 1 < \pi_i^c \leq q\}$; that is, $I_{iq} = 1$ if $i \in G_q$ and$I_{iq} = 0$ otherwise. These groups represents the implicit stratification. The reason for using this variable is justified using the entropy in Berger (2003b). A series of simulations in Berger (2003a, 2003b) shows that it is recommended to incorporate (10) in the working regression with systematic samples. An another approach proposed by Brewer (2002, page 159) consists on creating pseudo strata with at least two sampled units and assuming high entropy within strata. The approach are studied via simulation in Section 5.

4 Alternative Estimators for the Variance

If we substitute $\breve{c}_i$ by $n_h(n_h - 1)^{-1}$ $(i \in U_h)$ into (8), we get

$$\widehat{\sigma}^2_{swr} = \sum_{h=1}^{H} n_h (n_h - 1)^{-1} \sum_{i \in s_h} (\breve{y}_i - \widehat{B}^*_h)^2; \tag{11}$$

where $\widehat{B}^*_h = n_h^{-1} \sum_{i \in s_h} \breve{y}_i$. The variance estimator $\widehat{\sigma}^2_{swr}$ is the usual stratum-by-stratum Hansen-Hurwitz variance estimator. We note that (8) is as simple as (11) to compute, as both involve single sums and does not depend on unknown quantities. It is well known that variance estimation is greatly simplified by treating the sample as if the units were sampled with replacement. This approach is often adopted in practice. However this approach usually leads to overestimation of the variance. The Hájek estimator is as simple as the with-replacement variance estimator (11) and has a smaller bias.

If $\breve{c}_i = (1-\pi_i)\log(1-\pi_i)\pi_i^{-1}$, (8) is algebraically equivalent to the Rosén (1991) estimator implemented by Statistics Sweden. If $\breve{c}_i = (1-\pi_i)\ [1- d_h^{-2}\ \sum_{j\in s}(1-\pi_i)^2]^{-1}\ (i \in U_h)$, (8) gives the Deville (1999) variance estimator. These estimators are close to (8).

The Brewer's family of simple estimators also merited consideration (Brewer 2002, Chap. 9). This family uses the approximate formula for the π_{ij} derived by Hartley and Rao (1962). An estimator of this family is given by

$$\widehat{\sigma}^2_{Brewer} = \sum_{i\in s} \breve{c}_i^*\ \widehat{e}_i^{*2} \tag{12}$$

where $\widehat{e}_i^* = \breve{y}_i - \sum_{h=1}^H \widehat{B}_h^*\ z_{ih}$ are the ordinary least squares (OLS) residuals and $\widehat{B}_h^*$ is the OLS coefficient defined by

$$\widehat{B}_h^* = \left(\sum_{j\in s} z_{jh}^2\right)^{-1} \sum_{i\in s} \breve{y}_i\ z_{ih}$$

Note that $\widehat{B}_h^* = n_h^{-1}\sum_{i\in s_h}\breve{y}_i$. Brewer (2002, Chap. 9) proposed different choice for $\breve{c}_i^*\ (i \in U_h)$:

(i) $\breve{c}_i^* = 1-\pi_i$

(ii) $\breve{c}_i^* = n_h(n_h-1)^{-1}(1-\pi_i) = \breve{c}_i$

(iii) $\breve{c}_i^* = \breve{c}_i + (n_h-1)^{-1}\left[n_h^{-1}\sum_{j\in U}\pi_j^2 - \pi_i\right]$

The first choice ignores the correction of DF and is not recommended when few unit are sampled per stratum. The last choice depends on $\sum_{j\in U}\pi_j^2$ which is unknown if the π_i for $i \notin s$ are not available. With the second choice, the same weights are used in (8) and in (12). In this case, the only difference is in the regression coefficient: we have the WLS regression coefficients and in (8), and the OLS regression coefficients in (12). In the rest of this section, we show why WLS regression coefficients are recommended.

Let $\widehat{\mathbf{Z}} = \sum_{i\in s}(z_{i1},...,z_{iH})^T = (n_1,...,n_H)^T$ and $\mathbf{B} = (B_1,...,B_H)^T$. As $\widehat{\mathbf{Z}}$ is a fixed (non-random) vector, we have $\sigma^2_{Y^*} = var(\widehat{Y} - \widehat{\mathbf{Z}}\mathbf{B})$; where $var(\cdot)$ denotes the variance operator. Thus (2) equals

$$\sigma^2_{Y^*} = \sum_{h=1}^H \sum_{i\in U_h} \pi_i(1-\pi_i)(\breve{y}_i - B_h)^2 + \Delta \tag{13}$$

where B_h is defined by (6) and

$$\Delta = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} (\pi_{ij} - \pi_i\pi_j)\ e_i e_j$$

which is negligible compared to the first term of (13), as if we replace (4), we have $\Delta = 0$ (see also Brewer, 2002 Chap. 9). However, as the actual π_{ij} are different from (4), Δ approximately equals zero. Consider a family of approximation given by

$$\sigma^2 = \sum_{h=1}^{H} \sum_{i \in U_h} \pi_i (1 - \pi_i)(\breve{y}_i - \beta_h)^2$$

where β_h is any constant. It is well known that the β_h minimise σ^2 when $\beta_h = B_h$. Thus, the error $|\sigma^2 - \sigma_Y^2|$ of the approximation is minimal when $\beta_h = B_h$; that is, when $\sigma^2 = \sigma_{Haj}^2$. Finally, we recommend to use WLS regression coefficients B_h, as this should reduce the error in the approximation. Thus, when $\breve{c}_i^* = \breve{c}_i$, the bias of $\widehat{\sigma}_{Haj}^2$ should be smaller than the bias of $\widehat{\sigma}_{Brewer}^2$.

5 Simulations

In this Section, the Hájek variance estimator is studied through simulation studies. We consider a population of $N = 7000$ with a study variable y_i generated from the distribution of weekly household total expenditures estimated from the 98-99 UK Family Expenditure Survey. The total household expenditures are adjusted for the differing sizes and compositions of households (Department of Social Security, 2001). This study variable has a skewed distribution with a coefficient of skewness of 2.57 and coefficient of variation of 0.6.

With a linear model, we generate a size variable correlated with the study variable and with a coefficient of correlation of 0.6. We use the Dalenius & Hodges (1959) method to construct H strata according to the size variable. We consider a proportional allocation over the strata and a design with a within strata first-order inclusion probabilities proportional to the size variable. We will compare the bias and the accuracy of (8), (11) and (12) when $\breve{c}_i^* = \breve{c}_i$. We use the Chao (1982) sampling design to select units in each stratum since the π_{ij} can be computed exactly (Chao, 1982). This allows us to compare the distribution of $\widehat{\sigma}_{YG}^2$ with the distribution of $\widehat{\sigma}_{Haj}^2$.

To compare the performance of the variances estimators, we draw $M = 1000$ Chao samples to compute the empirical relative bias (RB) and the empirical root mean square errors (RMSE) for each variance estimator: (3), (8) and (11). The RB of a variance estimator $\widehat{\sigma}^2$ is given by $RB(\widehat{\sigma}^2) = 100 Bias(\widehat{\sigma}^2)\sigma^{-2}$; where

$$Bias(\widehat{\sigma}^2) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\sigma}_m^2 - \sigma^2$$

where σ^2 is the empirical variance of the π-estimator; that is, $\sigma^2 = M^{-1} \sum_{m=1}^{M} (\widehat{Y}_m - \overline{Y})^2$, $\overline{Y} = M^{-1} \sum_{m=1}^{M} \widehat{Y}_m$, $\widehat{Y}_m$ is the point estimates (1) and $\widehat{\sigma}_m^2$ is a variance estimates from the mth sample drawn. In addition, we present values of ratios of RMSE: $RMSE(\widehat{\sigma}_{Haj}^2)/RMSE(\widehat{\sigma}_{YG}^2)$, $RMSE(\widehat{\sigma}_{Brewer}^2)/RMSE(\widehat{\sigma}_{YG}^2)$

and $RMSE(\widehat{\sigma}^2_{swr})/RMSE(\widehat{\sigma}^2_{YG})$, where

$$RMSE(\widehat{\sigma}^2) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\sigma}^2_m - \sigma^2)^2}$$

These ratios are given in between bracket in column 4, 5 and 6. A ratio smaller than 1 means that the RMSE of the corresponding variance estimator is smaller than the RMSE of $\widehat{\sigma}^2_{YG}$.

Table 1 gives the empirical RB for different sampling fraction (f) and different number of strata (H). We expect $\widehat{\sigma}^2_{Haj}$ to be slightly biased as it is an estimator of an approximation (5) of the variance. Nevertheless, we see that the RB of $\widehat{\sigma}^2_{Haj}$ is negligible and as the same order of the RB of $\widehat{\sigma}^2_{YG}$, but not always. Although, $\widehat{\sigma}^2_{YG}$ is unbiased, the RB of $\widehat{\sigma}^2_{YG}$ is different from zero, as RB is an estimate of the actual bias. For small sampling fraction ($f \leq 0.05$), $\widehat{\sigma}^2_{Haj}$ is as accurate as $\widehat{\sigma}^2_{swr}$ and $\widehat{\sigma}^2_{swr}$ is even better when the number of strata is large. This is not surprising as the FPC is negligible for small sampling fractions. For large sampling fraction, $\widehat{\sigma}^2_{swr}$ has a large positive bias and $\widehat{\sigma}^2_{Haj}$ is a better option. As expected (see last paragraph of Section 4), the bias of the Brewer estimator $\widehat{\sigma}^2_{Brewer}$ is slightly larger than the bias $\widehat{\sigma}^2_{Haj}$.

The Hájek variance estimator is computationally simpler than the Sen-Yates-Grundy variance estimator and yet leads to values close to the Sen-Yates-Grundy variance. This conclusion is based on this simulation study and other studies may or may not confirm these results. However, Hájek (1964) derived (8) for the maximum entropy rejective sampling design. The high entropy of the Chao sampling design can explain why $\widehat{\sigma}^2_{Haj}$ is as good as $\widehat{\sigma}^2_{YG}$.

Table 2 gives the empirical RB and ratios RMSE when the sample is selected using the systematic sampling design with unequal probabilities. This series of simulation is based on the same data-set sorted according to the size variable. The systematic sample is selected assuming that the population is composed of a single stratum. However, for variance estimation, we create H pseudo strata. These pseudo strata are constructed as above. The column $\widehat{\sigma}^2_B$ gives the RB of Berger (2003a, 2003b) variance estimator (see Section 3.2) with a working regression having two independent variables: the intercept (as U is composed of a single stratum) and an additional independent variable (10). The RB of $\widehat{\sigma}^2_B$ is the same for varying values of H, as $\widehat{\sigma}^2_B$ does not depends on the pseudo-strata. The value in between brackets give ratios of RMSE; that is $RMSE(\widehat{\sigma}^2_{Haj})/RMSE(\widehat{\sigma}^2_B)$, $RMSE(\widehat{\sigma}^2_{swr})/RMSE(\widehat{\sigma}^2_B)$. We have intentionally omitted the RB of $\widehat{\sigma}^2_{YG}$ in Table 2. Although, it is possible to compute exactly the π_{ij} of the systematic sampling design (Connor, 1966; Pinciaro, 1978; Hidiroglou and Gray, 1980), most of π_{ij} equal zero, implying that $\widehat{\sigma}^2_{YG}$ is biased. Therefore, $\widehat{\sigma}^2_{YG}$ can be misleading and is not recommended for systematic sampling (Särndal *et al.*, 1992 p. 47).

f	H	$\widehat{\sigma}^2_{YG}$	$\widehat{\sigma}^2_{Haj}$	$\widehat{\sigma}^2_{Brewer}$	$\widehat{\sigma}^2_{swr}$
0.01	15	2.93	7.06 (1.14)	7.20 (1.14)	3.58 (1.00)
0.01	30	0.02	3.42 (1.56)	4.71 (1.56)	0.99 (1.00)
0.01	50	5.07	13.04 (1.56)	13.67 (1.56)	6.58 (1.00)
0.05	15	3.36	3.73 (1.02)	3.74 (1.02)	7.18 (1.00)
0.05	30	-4.72	-4.34 (1.06)	-4.33 (1.06)	-0.76 (1.00)
0.05	50	0.09	-0.10 (1.09)	-0.05 (1.09)	4.53 (1.00)
0.08	15	5.12	5.32 (1.01)	5.34 (1.01)	12.00 (1.01)
0.08	30	-1.02	-0.62 (1.03)	-0.60 (1.03)	4.88 (1.00)
0.08	50	1.32	1.04 (1.06)	1.06 (1.06)	8.84 (1.01)
0.10	15	3.44	3.69 (1.01)	3.72 (1.01)	12.11 (1.02)
0.10	30	-3.53	-3.12 (1.02)	-3.10 (1.02)	3.81 (1.00)
0.10	50	0.40	0.16 (1.04)	0.18 (1.04)	9.95 (1.02)
0.15	15	3.46	3.99 (1.01)	4.05 (1.01)	15.65 (1.01)
0.15	30	10.11	10.52 (1.03)	10.58 (1.03)	25.92 (1.05)
0.15	50	5.53	5.50 (1.03)	5.55 (1.03)	21.63 (1.14)
0.20	15	2.70	3.36 (1.00)	3.47 (1.00)	17.77 (1.00)
0.20	30	1.65	2.12 (1.01)	2.23 (1.01)	21.70 (1.06)
0.20	50	-4.53	-4.43 (1.02)	-4.34 (1.02)	16.07 (1.08)
0.25	15	-1.83	-0.55 (1.01)	-0.34 (1.01)	20.98 (1.03)
0.25	30	-1.40	-0.68 (1.01)	-0.51 (1.01)	22.16 (1.01)
0.25	50	5.41	6.18 (1.02)	6.34 (1.02)	33.17 (1.02)

Table 1: RB (%) of $\widehat{\sigma}^2_{Haj}$, $\widehat{\sigma}^2_{YG}$, $\widehat{\sigma}^2_{Brewer}$ and $\widehat{\sigma}^2_{swr}$ with Chao sampling. Ratios of RMSE are in between brackets.

f	H	$\widehat{\sigma}^2_{B}$	$\widehat{\sigma}^2_{Haj}$	$\widehat{\sigma}^2_{swr}$
0.05	2	70.27	64.62 (0.85)	68.18 (0.85)
0.05	5	70.27	90.36 (1.33)	95.70 (1.35)
0.05	12	70.27	29.91 (0.65)	38.27 (0.69)
0.10	5	39.01	18.59 (0.39)	23.29 (0.40)
0.10	10	39.01	2.16 (0.31)	7.61 (0.32)
0.10	25	39.01	49.25 (1.51)	63.70 (1.64)
0.20	10	10.33	11.71 (1.33)	18.73 (1.34)
0.20	20	10.33	6.45 (0.98)	14.90 (1.01)
0.20	50	10.33	-37.63 (0.32)	-27.02 (0.35)

Table 2: RB (%) of $\widehat{\sigma}^2_{B}$, $\widehat{\sigma}^2_{Haj}$ and $\widehat{\sigma}^2_{swr}$ with unequal probability systematic sampling. Ratios of RMSE are in between brackets.

As far as the RMSE is concerned, $\widehat{\sigma}^2_{Haj}$ sounds to be the best choice when the number of pseudo strata H is large, but not too large. When H is large, $\widehat{\sigma}^2_{Haj}$ can be worst than $\widehat{\sigma}^2_{B}$ and could even have a large negative bias, which is not

recommended for inference. Although $\widehat{\sigma}_B^2$ is not the most accurate estimator, it appears to be the most conservative choice, as its bias appears to be always positive.

6 Conclusion

Variance with unequal probability sampling without replacement can be easily estimated with the Hájek (1964) variance estimator. The contribution of this paper is to give an alternative expression of this estimator as a weighted sum of residuals. This alternative expression is computationally simpler than the Sen-Yates-Grundy variance estimator and does not require computation of joint-inclusion probabilities. Moreover, simulations show that the Hájek variance estimator is as accurate as the Sen-Yates-Grundy variance estimator.

Acknowledgements

The 98-99 UK Family Expenditure Survey data were made available by the Office for National Statistics (UK). This work was supported by the European research project entitled "Data Quality in Complex Surveys within the New European Information Society" (DACSEIS). The author is grateful to Chris Skinner (University of Southampton, UK), Dan Hedlin (University of Southampton, UK) and Pascal Rivière (INSEE, France) and to the referee for helpful comments.

References

ANDERSSON, C. AND NORDBERG, L. (1994), A Method for Variance Estimation of Non-Linear Function of Totals in Surveys. **Journal of Official Statistics**, 10, pp. 396-405.

BERGER, Y.G. (1998). Rate of Convergence for Asymptotic Variance for the Horvitz-Thompson Estimator. **Journal of Statistical Planning and Inference**, 74, pp. 149–168.

BERGER, Y.G. (2003a). A Modified Hájek Variance Estimator for Systematic Sampling. **Statistics in Transition**, Polish Statistical Association, June Issue.

BREWER, K.R.W. (2002). **Combined Survey Sampling Inference** (Weighing Basu's Elephants), Arnold publisheds.

BERGER, Y.G. (2003b). Variance Estimation for Systematic Sampling with Unequal Probabilities under Fixed Ordering of a Population. **SSRC working papers series, Southampton**, paper submitted.

CHAO, M.T. (1982). A General Purpose Unequal Probability Sampling Plan. **Biometrika**, 69,pp. 653–656.

CONNOR, W. S. (1966). An exact formula for the probability that two specified sample units will occur in a sample drawn with unequal probabilities and without replacement, **Journal of the American Statistical Association**, 61, pp. 384-390.

DALENIUS, T. & HODGES, J.L. (1959). Minimum Variance Stratification. **Journal of the American Statistical Association**, 54, pp. 88–101.

DEPARTMENT OF SOCIAL SECURITY (2001) Households Below Average Income *1999/00*, Appendix 2: Methodology.

DEVILLE, J.C. (1999) Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. **Survey Methodology**, 25, pp. 193–203.

GABLER, S. (1984). On Unequal Probability Sampling: Sufficient Conditions for the Superiority of Sampling without Replacement. **Biometrika**, 71, pp. 171–175.

HÁJEK, J. (1964). Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. **Annal of Mathematical Statistics**, 35, pp. 1491–1523.

HÁJEK, J. (1981). **Sampling from a Finite Population** (New York, Marcel Dekker).

HANSEN, M.H & HURWITZ, W.N. (1943). On the Theory of Sampling from Finite Population. **Annal of Mathematical Statistics**, 14, pp. 333–362.

HARTLEY, H.O. & RAO, J.N.K. (1962). Sampling with Unequal Probabilities and without Replacement. **Annal of Mathematical Statistics**, 33, pp. 350–374.

HIDIROGLOU, M.A. & GRAY, G.B. (1980), Construction of joint probability of selection for systematic pps sampling, **Applied Statistics**, 29, pp. 107 112.

HORVITZ, D.G. & THOMPSON, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. **Journal of the American Statistical Association**, 47, pp. 663–685.

NARAIN, R.D. (1951). On Sampling without Replacement with Varying Probabilities, **Journal of the Indian Society of Agricultural Statistics**, 3, pp. 169–174.

PINCIARO, S.J. (1978). An algorithm for calculating joint inclusion probabilities under PPS systematic sampling, **ASA Proceedings of Survey Research Methods Section**, pp. 740.

RAO, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling without Replacement. **Journal of the Indian Statistical Association**, 3, pp. 173–180.

ROSÉN, B. (1991). Variance for systematic pps-sampling. **Report 1991:15, Statistics Sweeden**.

SAMPFORD, M.R. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. **Biometrika**, 54, pp. 494–513.

SÄRNDAL, C.E., B. SWENSON & J. H. WRETMAN (1992). **Model Assisted Survey Sampling**, Springer-Verlag.

SEN, P.K. (1953). On the Estimate of the Variance in Sampling with Varying Probabilities. **Journal of the Indian Society of Agricultural Statistics**, 5, pp. 119-127.

SKINNER, C.J. (1989). **In:** SKINNER, C.J., HOLT, D. AND SMITH, T.M.F. (Ed.), **Analysis of Complex Surveys** (Chichester, Wiley).

SMITH, T.M.F. (2001). Biometrika Centenary: Sample Surveys. **Biometrika**, 88, pp. 167–194.

YATES, F. & GRUNDY, P. M. (1953). Selection without Replacement from within Strata with Probability Proportional to Size. **Journal of the Royal Statistical Society Serie B**, 1, pp. 253–261.