COMPARISON OF VARIANCE ESTIMATORS UNDER RAO-SAMPFORD METHOD: A SIMULATION STUDY

David Haziza, Fulvia Mecatti* and Jon N.K. Rao
Statistics Canada, University of Milan-Bicocca and Carleton University

KEY WORDS: Approximate joint inclusion probabilities; Efficiency; Relative bias; Unequal probability sampling without replacement.

1. The problem

Unequal probability sampling with inclusion probabilities π_i exactly proportional to a measure of size x, known for each unit (often called πPS) is extensively used in large-scale surveys, especially for the selection of primary sampling units in multi-stage sampling. For simplicity, we focus on uni-stage sampling from a finite population U of size N. The Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i\in s} y_i/\pi_i$, with variance $V\left(\hat{Y}_{HT}\right)$, is used to estimate the total $Y = \sum_{i\in U} y_i$ of a characteristic of interest y, which is approximately proportional to x, where s denotes a sample of fixed size n and $\pi_i = n x_i / X$ with $X = \sum_{i\in U} x_i$. The well known Sen-Yates-Grundy variance estimator

$$v_{SYG}\left(\hat{Y}_{HT}\right) \equiv v_{SYG} = \sum_{i\in s}\sum_{j<i\in s} \frac{\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad (1)$$

is exactly unbiased. On the other hand, it involves the joint inclusion probabilities π_{ij}. Consequently, several alternative variance estimators, based on approximating π_{ij} in terms of the π_i's only, have been proposed; see Section 2. It is also often stated that the exactly unbiased variance estimator v_{SYG} can be very unstable.

The present study has been performed under Rao-Sampford πPS sampling without replacement (Rao, 1965; Sampford, 1967). This method has several desirable properties: first order inclusion probabilities π_i are exactly proportional to size x_i, v_{SYG} is nonnegative for any sample size n and the variance of $\hat{Y}_{HT}$ is uniformly smaller than the corresponding variance under probability proportional to size sampling with replacement (*PPS*). Moreover, joint inclusion probabilities π_{ij} can be exactly calculated by recursive formulae already implemented in SAS. Hence v_{SYG} can be readily calculated for any n using SAS software. The main purpose of this paper is to use simulations in order to first study the relative biases of the approximate variance estimators, then to select those that perform well in terms of bias and to compare them with the exact variance estimator v_{SYG} in terms of relative efficiency.

2. Approximate variance estimators

Our study deals with a collection of 12 approximate variance estimators. The first proposal of approximating the joint inclusion probabilities π_{ij} in terms of first order inclusion probabilities π_i only, is the well-known asymptotic first order approximation by Hartley and Rao (1962) for fixed sample size $n \geq 2$ and increasing population size $N \to \infty$, under the randomized systematic πPS sampling. It may be noted that the exact evaluation of π_{ij}'s for this design is cumbersome as n increases, unlike the Rao-Sampford design.

Substituting this first order approximation in (1) leads to the approximate variance estimator

$$v_{HR} = \frac{1}{n-1}\sum_{i\in s}\sum_{j<i\in s}\left(1-\pi_i-\pi_j+\frac{1}{n}\sum_{i\in U}\pi_i^2\right) \times\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \qquad (2)$$

Asok and Sukhatme (1976) showed that Hartely-Rao's first order approximation works under the Rao-Sampford design also. They also derived a second order approximation of π_{ij} under the Rao-Sampford design. Substituting this second order approximation in (1) provides a further approximate variance estimator, strictly related to the Rao-Sampford method, given by

*Department of Statistics, University of Milan-Bicocca Via Bicocca degli Arcimboldi, 8, Ed. U7, 20126 Milan, Italy. Fulvia.mecatti@unimib.it

$$v_{AS} = \frac{1}{n-1}\sum_{i\in s}\sum_{j<i\in s}\left[1 + n\left(A_{ij}^{-1} - 1\right)\right]\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \quad (3)$$

where

$$A_{ij} = 1 + \frac{\pi_i + \pi_j}{n} - \frac{1}{n^2}\sum_{i\in U}\pi_i^2 + \frac{2}{n^2}\left(\pi_i^2 + \pi_j^2\right) - \frac{2}{n^3}\sum_{i\in U}\pi_i^3 - \frac{n-2}{n^2}\pi_i\pi_j + \frac{n-3}{n^3}\left(\pi_i + \pi_j\right)\sum_{i\in U}\pi_i^2 - \frac{n-3}{n^4}\left(\sum_{i\in U}\pi_i^2\right)^2.$$

A different asymptotic framework has been considered by Hájek (1964) by assuming $\sum_{i\in U}\pi_i(1-\pi_i)\to\infty$, which implies both $n\to\infty$ and $(N-n)\to\infty$. Hàjek's approximation for π_{ij} leads to a different approximate variance estimator, v_H. Still based on Hájek results is the estimator proposed by Berger (1998), v_{BH}. Note that v_H and v_{BH} are asymptotically equivalent. It can also be shown (Berger, 1996) that the Rao-Sampford design is *close* to the optimum fixed size πPS design under a maximum entropy argument. We next consider three additional variance estimators related to maximum entropy sampling plans: two estimators by Deville (1999), v_{D1} and v_{D2}, and one proposed by Rosen (1991), v_R. We have also examined a family of approximate variance estimators proposed recently (Brewer, 2002; Brewer and Donadio, 2003); particularly four estimators v_{Bk} $k=1,...,4$, arising from four heuristic approximations to π_{ij}. To the same family we finally added a very simple approximation, v_{BM}, which follows by comparing Hartley-Rao's and Asok-Sukhatme's asymptotics with Brewer's approximation. The π_{ij} approximation involved in v_{BM} also coincides with the starting trial value suggested by Brewer (Brewer and Donadio, 2003).

All the variance estimators we are dealing with and valid for any πPS sampling plan, can be expressed in the following common form

$$v\left(\hat{Y}_{HT}\right) = \sum_{i\in s} c_i\left(\frac{y_i}{\pi_i} - \hat{B}\right)^2, \quad (4)$$

where $\hat{B} = \left(\sum_{i\in s} a_i\right)^{-1}\sum_{i\in s} a_i\left(y_i/\pi_i\right)$ and a_i and c_i are fixed weights to be chosen, for $i\in s$, as function of first order inclusion probabilities only. Note that by letting $\hat{e}_i = y_i/\pi_i - \hat{B}$, $\hat{B}$ can be interpreted as the estimated weighted regression coefficient of the *ratio* model $y_i/\pi_i = B + e_i$ using a_i as the weights, where e_i are uncorrelated random variables with zero mean . Also note that (4) equals zero when $y_i \propto x_i$.

First letting $a_i = 1$, so that $\hat{B} = \hat{Y}_{HT}/n$, we have:

1. $c_i = \left(1 - \pi_i - n^{-1}\sum_{i\in s}\pi_i + n^{-1}\sum_{i\in U}\pi_i^2\right)n/(n-1)$ provides v_{HR} (see the Appendix);
2. $c_i = (1-\pi_i)\left[1 - \sum_{i\in s} d_i^2\right]^{-1}$, where $d_i = (1-\pi_i)/\sum_{i\in s}(1-\pi_i)$ gives v_{D2}.

Moreover, the estimators in Brewer's family are obtained by choosing $c_i = b_i^{-1} - \pi_i$ with the following values of b_i:

3. $b_i = (n-1)/(n-\pi_i) = (1-\pi_i/n)^{-1}(n-1)/n$ gives v_{B1};
4. $b_i = \left(1 - n^{-2}\sum_{i\in U}\pi_i^2\right)^{-1}(n-1)/n$ yields v_{B2};
5. $b_i = \left(1 - n^{-1}2\pi_i + n^{-2}\sum_{i\in U}\pi_i^2\right)^{-1}(n-1)/n$ leads to v_{B3};
6. $b_i = \left\{1 - n^{-1}(n-1)^{-1}\left[(2n-1)\pi_i - \sum_{i\in U}\pi_i^2\right]\right\}^{-1}(n-1)/n$ results in v_{B4}. Furthermore, the simplest choice:
7. $b_i = (n-1)/n$ gives v_{BM}.

By putting $a_i = c_i \neq 1$ we have:

8. $c_i = a_i = (1-\pi_i)n/(n-1)$ leads to v_H;
9. $c_i = a_i = A(1-\pi_i)n/(n-1)$ where $A = \sum_{i\in s}(1-\pi_i)/\sum_{i\in U}\pi_i(1-\pi_i)$, gives v_{BH};
10. $c_i = a_i = (1-\pi_i)\left[1 - \sum_{i\in s} d_i^2\right]^{-1}$ provides v_{D1}.

Finally, for $a_i \neq 1 \neq c_i$ the choice:

11. $c_i = (1-\pi_i)n/(n-1)$ and $a_i = (1-\pi_i)[log(1-\pi_i)]/\pi_i$ yields v_R.

Previous empirical studies are included in Stehman and Overton (1994) and Berger (1996) focusing mainly on the approximation of joint inclusion probabilities. More recently, Matei and Tillé (2003) and Brewer and Donadio (2003) considered larger sets of approximate variance estimators, mostly concentrating on the part played by the joint inclusion probabilities into the variance estimation process. In the present study, the collection of variance estimators analyzed coincides only partly with those sets and the array of populations explored by varying key simulation parameters has been enlarged. Besides, we focus on the Rao-Samford design which is a simpler practical choice with respect to the πPS sampling plans considered in previous studies, and *close* to the maximum entropy fixed size plan.

3. The simulation study

Our main purpose is to investigate the performances of the twelve approximate variance estimators presented in Section 2, in terms of bias and efficiency with respect to the exactly unbiased Sen-Yates-Grundy estimator (1).

3.1 Implementation

In order to explore a scenario sufficiently varied accordingly to the simulations goals, we started by considering four *natural* populations with the characteristics reported in Table 1, namely the variability of both the survey variable and the size variable as measured by their coefficients of variation $cv(y)$ and $cv(x)$, the population size N and the correlation ρ between x and y. Those are assumed as *key parameters* for the simulation.

Figure 1 also shows the areas reviewed by considering three levels of variability in x and y (low, medium and high) as well as the areas not explored since a πPS design is an appropriate practical option only if both the following conditions hold:

i) the variability of the size variable x is *sufficiently high* for πPS to differ significantly from simple random sampling;
ii) a positive relationship between y and x can be assumed.

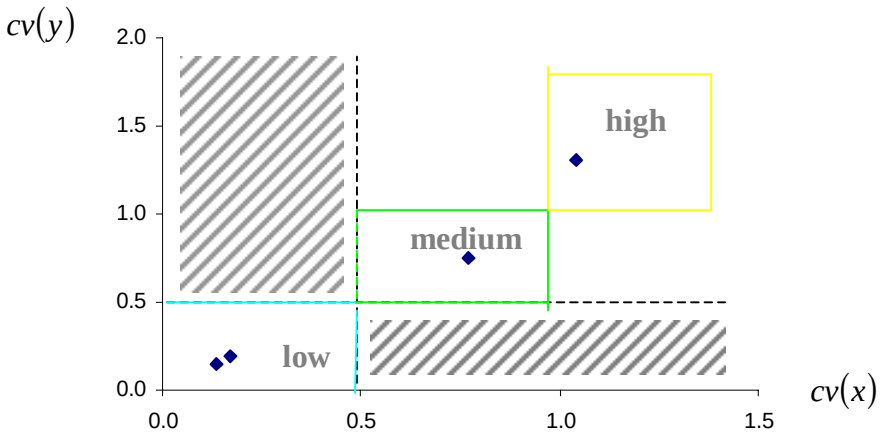
Besides variability in x and y, artificial populations have been generated by varying the population size ($N = 20, 50, 100$), the correlation level ($\rho = 0.6,\ 0.9$) and the sample fraction $f = n/N$, $(f = 0.1, 0.2, \cdots, 0.7)$, the latter under the constraint $\pi_i < 1,\ \forall i \in U$, leading to $f \geq 0.2$ for smaller populations only.
The artificial populations were generated accordingly to the model: $y_i = \gamma x_i + \varepsilon_i$ in order to simulate the required approximated proportionality between y and x, where ε_i are uncorrelated normal random variables with zero mean. More precisely, $x_i,\ \forall i \in U$, are finite realizations of a gamma distribution with expectation equal to 100 while y-values are produced conditionally given x_i as $y_i | x_i = 5x_i + N(0, \sigma^2 x_i)$.
Simulations given with 50,000 runs have been performed to ensure a Monte Carlo error always negligible. Since SAS software provides exact joint inclusion probabilities for sampled pairs of units $(i \neq j) \in s$ only, for each population considered we first produced a Monte Carlo approximation for the actual variance $V(\hat{Y}_{HT})$ by drawing 250,000 samples.

Table 1: Characteristics of the 4 natural populations considered

	KISH (1965, p. 42)	**COCHRAN 1** (1963, p. 204)	**COCHRAN 2** (1963, p. 325)	**SUKHATME** (1970, P. 185)
$cv(x)$	1.04	0.17	0.14	0.77
$cv(y)$	1.31	0.19	0.15	0.76
N	20	10	10	34
ρ	0.74	0.97	0.65	0.93

The four *natural* populations can be represented as points in the plane with $cv(x)$ on the horizontal axis and $cv(y)$ on the vertical axis as shown in Figure 1. In the same plane we generated more than 20 *artificial* populations in order to enlarge the scenarios explored.

Figure 1: 4 natural populations and areas simulated by generating artificial populations



3.2 Main results

First regarding bias, simulation results for each approximate variance estimator, generically denoted by *v*, have been summarized in terms of relative bias percentage (*RB*)

$$RB = \frac{E(v) - V(\hat{Y}_{HT})}{V(\hat{Y}_{HT})} \times 100.$$

Focusing first on low variability populations, located in the lower left of Figure 1, except for v_{D2} and v_{BM} which can be heavily biased with $10\% < |RB| < 30\%$ (and even larger for larger sample fraction *f*), all the remaining estimators appear approximately unbiased having $|RB| < 5\%$. Among those, estimators v_{AS} and v_{B4} are the best performers with $|RB| < 1\%$ especially when *n* is small compared to *N*. On the other hand, estimators related to Hájek approximation and maximum entropy design, v_H, v_{BH}, v_{D1} and v_R along with two included in the Brewer's family v_{B1} and v_{B2}, seem sensitive to the correlation between *x* and *y*, rapidly turning biased as ρ decreases.

The picture is similar for medium variability populations, located around the centre of Figure 1. Here, v_{BM} is still biased ($1\% < |RB| < 10\%$) but performs noticeably better than v_{D2}; estimators v_{AS} and v_{B4} are still essentially unbiased with $|RB| < 1\%$. To them we may add v_{BH} and v_{HR} when $N \geq 50$.

Populations highly dispersed, located in the upper right area of Figure 1, represent our worst scenario, scarcely examined in former empirical work. Here, the two best performers in previous cases, v_{AS} and v_{B4}, again emerge by retaining $|RB| < 1\%$ over all the populations simulated. Estimators v_{BH} and v_{B3} also are approximately unbiased for larger *N*. Surprisingly, for the difficult case of high variability in *x* and *y* also the simplest approximation v_{BM} performs well, essentially depending upon greater values of *N*, *f* and ρ.

The best performers in terms of bias were then compared with the exactly unbiased variance estimator v_{SYG} in terms of relative efficiency as measured by the customary ratio (Eff_v)

$$Eff_v = \frac{V(v_{SYG})}{MSE(v)}.$$

Table 2 reports a synthesis of the general tendency in the three cases of low, medium and high variability illustrated in Figure 1. Cells in Table 2 are filled with the simulated ratio Eff_v only if the correspondent variance estimator is approximately unbiased.

Notice that the efficiency ratio is close to unity over all values in Table 2, indicating the approximate variance estimators performs similar to the exact unbiased estimator in term of stability. Besides, the efficiency ratio is never less than 1; hence we may conclude that when the approximate variance estimator is approximately unbiased it is never less efficient than the exact estimator. Moreover, some percentage gains in efficiency might be further pursued. Between the two best performers v_{AS} and v_{B4}, while the former appears

as efficient as the Sen-Yates-Grundy estimator, the Brewer estimator v_{B4} is 12% more efficient. The higher the population variability, the more might be the efficiency gain: for law variability populations the possible gain by using an approximately unbiased variance estimator are about 3.5% which can increase to 13% for medium variability, and up to 20% for high variability populations. Note that the greater efficiency gains correspond to the simplest approximate variance estimator v_{BM} .

Table 2: Efficiency ratio Eff_v for the approximately unbiased variance estimators in the three cases of low, medium and high variability of y and x

	low	medium	high
v_{AS}	1.0005	1.0061	1.0327
v_{B4}	**1.0197**	**1.0594**	**1.1231**
v_{BH}	1.0041	1.0174	1.0492
v_{HR}	1.0044	1.0262	
v_{BM}			**1.1972**
v_{H}	**1.0359**	1.1035	
v_{D1}	**1.0340**	1.0923	
v_{R}	**1.0356**	1.1023	
v_{B1}	**1.0344**	1.0984	
v_{B2}	1.0278	**1.1345**	
v_{B3}	1.0212	1.0633	1.1317

APPENDIX

Estimator v_{HR} as originally defined by (2) involves a double summation. Since it does not depend upon the joint inclusion probabilities, it can be rewritten in term of single sums only as follows

$$v_{HR} = \frac{1}{n-1}\sum_{i\in s}\sum_{j<i\in s}\left(1-\pi_i-\pi_j+n^{-1}\sum_{i\in U}\pi_i^2\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

$$= \frac{1}{n-1}\left[\left(1+n^{-1}\sum_{i\in U}\pi_i^2\right)\sum_{i\in s}\sum_{j<i\in s}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 - 2\sum_{i\in s}\sum_{j<i\in s}\pi_i\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\right]$$

$$= \frac{1}{n-1}\left[\left(1+n^{-1}\sum_{i\in U}\pi_i^2\right)n\sum_{i\in s}\left(\frac{y_i}{\pi_i}-\frac{\hat{Y}_{HT}}{n}\right)^2 - n\sum_{i\in s}\left(\pi_i+\ n^{-1}\sum_{i\in s}\pi_i\right)\left(\frac{y_i}{\pi_i}-\frac{\hat{Y}_{HT}}{n}\right)^2\right]$$

$$= \frac{n}{n-1}\sum_{i\in s}\left(1-\pi_i-\ n^{-1}\sum_{i\in s}\pi_i\ +\ n^{-1}\sum_{i\in U}\pi_i^2\right)\left(\frac{y_i}{\pi_i}-\frac{\hat{Y}_{HT}}{n}\right)^2$$

REFERENCES

Asok, C. and Sukhatme, B.V. (1976), "On Sampford's Procedure of Unequal Probability Sampling Without Replacement.", *Journal of the American Statistical Associatio*, **71**, 912-918.

Berger, Y.G. (1996), "On Sampling with Unequal probabilities Close to Rejective Sampling.", SSC Annual Meeting, Proceedings of the Survey Methods Section, 97-102.

Berger, Y.G. (1998), "Variance estimation using list sequential scheme for unequal probability sampling.", Journal of Official Statistics 14, 315-323.

Brewer, K.R.W. (2002), Combined Survey Sampling Inference, Weighing Basu's Elephants, Arnold, London.

Brewer, K.R.W. and Donadio, M.E. (2003), "The High Entropy Variance of the Horwitz-Thompson Estimator.", Survey Methodology 29, 189-196.

Deville, J. (1999), "Variance Estimation for Complex Statistics and Estimators : Linearization and Residual Techniques.", Survey Methodology 25, 193-203.

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

Hartley, H.O. and Rao, J.N.K. (1962), "Sampling with Unequal Probabilities and Without Replacement.", The Annals of Mathematical Statistics 33, 350-374.

Matei, A. and Tillé, Y. (2003), "Evaluation of variance estimators in unequal probability sampling.", Manuscript paper.

Rao, J.N.K. (1965), "On Two Simple Schemes of Unequal Probability Sampling Without Replacement.", Journal of the Indian Statistical Association 3, 173-180.

Rosen, B. (1991), "Variance estimation for systematic pps-sampling.", Report 1991:15, Statistics Sweden.

Sampford, M.R. (1967), "On Sampling Without Replacement with Unequal Probabilities of Selection.", Biometrika 54, 499-513.

Stehman, S.V. and Overton, S.W. (1994), "Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling.", Journal of the American Statistical Association 89, 30-43.