Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 315–323

Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling

Yves G. Berger[1]

The problem of variance estimation is discussed in the light of the list sequential scheme proposed by Chao (1982), in which units are selected without replacement and with unequal probabilities. The variance is hard to estimate as it requires a large number of second-order inclusion probabilities. We prove that it is unnecessary to compute all these probabilities. We show that variance estimation needs only N numbers, where N is the population size.

Key words: Variance estimation; sampling without replacement; Horvitz-Thompson estimator; Yates-Grundy estimator; inclusion probabilities; probability proportional-to-size sampling.

1. Introduction

Consider a finite population U_N consisting of N units labelled $i = 1, \ldots, N$. The population size N may be unknown. Let U_k be the subset of U_N comprising the k first units $\{1, \ldots, k\}$. We will use the Horvitz-Thompson estimator (1951) to estimate the population total $T_{(k)} = Y_1 + Y_2 + \ldots + Y_k$. This estimator is given by

$$\hat{T}_{(k)} = \sum_{i \in S_k} \frac{Y_i}{\pi_{(k;i)}} \tag{1}$$

where S_k is a sample of U_k. We assume that the size of S_k is constant and equal to n for all $k \geq n$. $\pi_{(k;i)}$'s $(i = 1, \ldots, k)$ denote the first-order inclusion probabilities for a population U_k.

Given that the size of the sample is fixed, a variance estimator of (1) is given by the Yates-Grundy estimator (1953):

$$\hat{V}_{(k)} = \sum_{j \in S_k} \sum_{\substack{i \in S_k \\ i<j}} \breve{\Delta}_{(k;i,j)} \left[\frac{Y_i}{\pi_{(k;i)}} - \frac{Y_j}{\pi_{(k;j)}} \right]^2 \tag{2}$$

where

$$\breve{\Delta}_{(k;i,j)} = \frac{\pi_{(k;i)}\pi_{(k;j)}}{\pi_{(k;i,j)}} - 1, \quad i < j \leq k \tag{3}$$

[1] Université Libre de Bruxelles, Laboratoire de Méthodologie du Traitement des Données, LB190, CP124, 44, Avenue Jeanne, B1050 Brussels, Belgium. E-mail: yvberger@ulb.ac.be.

Acknowledgments: The author wishes to thank the referees who provided a number of constructive comments that led to considerable improvement.

$\pi_{(k;i,j)}$ $(i,j = 1,\ldots,k)$ denotes the second-order inclusion probability of units i and j. These probabilities are dependent on the sampling scheme used. We call $\breve{\Delta}_{(k;i,j)}$'s the weights of the variance estimator (2). Let $\breve{\Delta}_{(k;,.)}$ be the matrix of these weights. The large number of weights involves heavy calculations for the Yates-Grundy estimator.

In Section 2, we define the probability proportional-to-size sampling. In Section 3, we introduce the Chao sampling scheme as well as one result linked to first-order inclusion probabilities. Section 4 gives major results concerning second-order inclusion probabilities. In Section 5, we analyze $\breve{\Delta}_{(k;,.)}$ and the variance estimator. A numerical example is provided in Section 6.

2. Probability Proportional-to-size Sampling

With a probability proportional-to-size sampling, the first-order inclusion probabilities satisfy

$$\pi_{(k;i)} = \frac{nX_i}{\sum_{j\in U_k} X_j} \tag{4}$$

where X_i's are values of some auxiliary variable. We simply assume that

$$0 < X_i \leq \frac{1}{n}\sum_{j\in U_k} X_j \tag{5}$$

for all $k > n$ and for all $i \leq k$. This means that the first-order inclusion probabilities $\pi_{(k;i)}$ are strictly proportional to X_i for all $k > n$. This hypothesis is more likely to break down when k is small, i.e., close to n. We can solve this problem by assuming that the values of the auxiliary variable show little dispersion for units occurring at the beginning of the population unit sequence.

When $\pi_{(k;i)}$ is exactly proportional to Y_i, the variance of the Horvitz-Thompson estimator becomes zero. For that reason, we prefer to implement a probability proportional-to-size sampling scheme.

3. Chao's Unequal Probability Sampling Scheme

Chao (1982) proposes a probability proportional-to-size sampling. This is a generalization of the McLeod and Bellhouse (1983) sampling scheme. Chao's sampling scheme provides the advantage of being sequential. In fact, the sample is selected through a simple sequential run of the population.

The sampling process of Chao (1982) allows us to pass from a sample S_k selected with inclusion probabilities $\pi_{(k;i)}$'s to a sample S_{k+1} selected with inclusion probabilities $\pi_{(k+1;i)}$'s. For this scheme, we simply draw the $(k+1)$th unit with the probability

$$w_k = \pi_{(k+1;k+1)}$$

If the $(k+1)$th unit is not drawn, then we take $S_{k+1} = S_k$; otherwise, we take $S_{k+1} = S_k \cup \{k+1\}\backslash\{j\}$; where j is a unit selected at random within S_k. This procedure starts

from an initial sample $S_n = U_n$ comprising the n first units of the population. This method is repeated until $k = N$.

The following lemma provides a relation between the first-order inclusion probability $\pi_{(k;i)}$ of the ith unit of U_k and the first-order inclusion probability $\pi_{(k+1;i)}$ of the ith unit of U_{k+1}.

Lemma 1.

$$\pi_{(k+1;i)} = \begin{cases} Q_{(k;i)}\pi_{(k;i)} & , \text{ if } i < k+1 \\ w_k & , \text{ if } i = k+1 \end{cases} \tag{6}$$

where

$$Q_{(k;i)} = 1 - w_k R_{(k;i)} \tag{7}$$

$$R_{(k;i)} = \begin{cases} \frac{1-\pi_{(n+1;i)}}{w_n} & \text{for} \quad k = n \\ \frac{1}{n} & \text{for} \quad k \geq n+1 \end{cases} \tag{8}$$

$$w_k = \pi_{(k+1;k+1)}$$

The proof of this lemma can be found in Appendix I.

4. Second-order Inclusion Probabilities

The second-order inclusion probabilities for Chao's sampling scheme can be calculated iteratively using the following theorem:

Theorem 1. *If* $i < j$,

$$\pi_{(k+1;i,j)} = \begin{cases} [Q_{(k;i)} + Q_{(k;j)} - 1]\pi_{(k;i,j)} & , \text{ if } \quad j < k+1 \\ w_k[1 - R_{(k;i)}]\pi_{(k;i)} & , \text{ if } \quad j = k+1 \end{cases} \tag{9}$$

The proof of this theorem can be found in Chao (1982, Lemma 2).

Bethlehem and Schuerhoff (1984) give a necessary and sufficient condition for the second-order inclusion probabilities to be strictly positive for a population U_k:

$$\#\{i : i \leq \ell \text{ and } \pi_{(\ell;i)} = 1\} \leq n - 1, \quad \text{for } \ell \text{ such that } n < \ell \leq k$$

By using (5), $\pi_{(\ell;i)} < 1$ for all i and ℓ such that $i \leq \ell \leq k$. Then this condition is always met. Therefore, within the framework of this article, we will never have zero second-order inclusion probabilities. Thus the Yates-Grundy estimator (2) is always unbiased.

Moreover, Chao (1982) states that the weights $\breve{\Delta}_{(k;i,j)}$ are always positive if Chao's sampling scheme is implemented. This ensures that the Yates-Grundy estimator will never be negative.

5. Variance Estimator

To compute the Yates-Grundy estimator for a population U_{k+1}, we must know the matrix $\breve{\Delta}_{(k+1;.,.)}$.

From Lemma 1 and Theorem 1, we derive a recursive relation for $\breve{\Delta}_{(k+1;i,j)}$.

Lemma 2. *If* $i < j$,

$$\breve{\Delta}_{(k+1;i,j)} = \begin{cases} \dfrac{Q_{(k;i)}Q_{(k;j)}}{Q_{(k;i)}+Q_{(k;j)}-1}\left[\breve{\Delta}_{(k;i,j)}+1\right]-1 & \text{if} \quad j < k+1 \\ \dfrac{Q_{(k;i)}}{1-R_{(k;i)}}-1 & \text{if} \quad j = k+1 \end{cases}$$

The following theorem shows that most elements of a given column of $\breve{\Delta}_{(k+1;.,.)}$ are identical. Theorem 2 is the corner stone of the present article.

Theorem 2. *If* $i < j$,

$$\breve{\Delta}_{(k+1;i,j)} = \begin{cases} \beta_{(k+1;i,j)} & \text{if} \quad j \le n+1 \\ \alpha_{(k+1;j)} & \text{if} \quad j > n+1 \end{cases}$$

where

$$\beta_{(k+1;i,j)} = -1 + \frac{\pi_{(n+1;i)}\pi_{(n+1;j)}}{\pi_{(n+1;i)}+\pi_{(n+1;j)}-1}\prod_{\ell=n+1}^{k} p_\ell \tag{10}$$

$$\alpha_{(k+1;j)} = -1 + \frac{n-w_{j-1}}{n-1}\prod_{\ell=j}^{k} p_\ell \tag{11}$$

$$p_\ell = \frac{\left(1-\dfrac{w_\ell}{n}\right)^2}{1-2\dfrac{w_\ell}{n}} \tag{12}$$

The proof of this theorem can be found in Appendix II.

We note that the vector $\alpha_{(k+1;.)}$ and some elements of the matrix $\beta_{(k+1;.,.)}$ are required to compute $\breve{\Delta}_{(k+1;.,.)}$. Using Theorem 2, we only need w_ℓ $(\ell = n+1, \ldots, k)$ and $\pi_{(n+1;i)}$ $(i = 1, \ldots, n+1)$ to compute the vector $\alpha_{(k+1;.)}$ and the matrix $\beta_{(k+1;.,.)}$. Therefore, it is only necessary to compute the following vector of length k:

$$\left\{\pi_{(n+1;1)},\ \pi_{(n+1;2)}, \ldots,\ \pi_{(n+1;n+1)},\ \pi_{(n+2;n+2)},\ \pi_{(n+3;n+3)}, \ldots,\ \pi_{(k;k)}\right\} \tag{13}$$

Moreover, using (4), it is clear that

$$\pi_{(n+1;i)} = \frac{nX_i}{C_{n+1}} \quad \text{if } i \le n+1$$

$$\pi_{(i;i)} = \frac{nX_i}{C_i} \quad \text{if } i > n+1$$

Where C_i's are cumulative totals of the auxiliary variable:

$$C_i = \begin{cases} C_{i-1}+X_i & \text{if } i > n \\ \displaystyle\sum_{\ell=1}^{n} X_\ell & \text{if } i = n \end{cases}$$

Thus the vector (13) needs only the C_i's, which can be computed during the sampling process. Finally, if we know the cumulative totals, it is rather easy to compute the matrix $\breve{\Delta}_{(k+1;.,.)}$.

The special structure of $\breve{\Delta}_{(k+1;.,.)}$ provides simplification for the Yates-Grundy estimator. Indeed, using

$$\frac{Y_i}{\pi_{(k+1;i)}} = \frac{C_{k+1}}{n}\frac{Y_i}{X_i}$$

and Theorem 2, the Yates-Grundy estimator becomes

$$\hat{V}_{(k+1)} = \left(\frac{C_{k+1}}{n}\right)^2 \left\{ \sum_{\substack{j \in S_{k+1} \\ j > n+1}} \alpha_{(k+1;j)} \sum_{\substack{i \in S_{k+1} \\ i<j}} D_{ij} + \sum_{\substack{j \in S_{k+1} \\ j \le n+1}} \sum_{\substack{i \in S_{k+1} \\ i<j}} \beta_{(k+1;i,j)} D_{ij} \right\} \tag{14}$$

where

$$D_{ij} = \left(\frac{Y_i}{X_i} - \frac{Y_j}{X_j}\right)^2$$

If we assume that the value of the auxiliary variable is small for units occurring at the beginning of the population unit sequence, there is no more than a small probability of selecting those units. This means that the first sum contains most of the terms. Fortunately, the first sum is the simplest to compute. It is therefore preferable to place the units having small value of the auxiliary variable at the beginning of the population unit sequence.

6. Numerical Example

Suppose that we want to select three units from a population U_{10}, composed of 10 units. Consider the values of the auxiliary variable:

{45, 30, 28, 40, 24, 49, 17, 62, 56, 29}.

The respective first-order inclusion probabilities are

i	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$
$\pi_{(10;i)}$	0.355	0.237	0.221	0.316	0.189	0.387	0.134	0.489	0.442	0.229

Using Theorem 2, we obtain the matrix $\breve{\Delta}_{(10;.,.)}$:

i	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$
1	0.319	0.328	0.288	0.589	0.311	0.563	0.237	0.269	0.385
2	–	1.171	0.435	0.589	0.311	0.563	0.237	0.269	0.385
3	–	–	0.471	0.589	0.311	0.563	0.237	0.269	0.385
4	–	–	–	0.589	0.311	0.563	0.237	0.269	0.385
5	–	–	–	–	0.311	0.563	0.237	0.269	0.385
6	–	–	–	–	–	0.563	0.237	0.269	0.385
7	–	–	–	–	–	–	0.237	0.269	0.385
8	–	–	–	–	–	–	–	0.269	0.385
9	–	–	–	–	–	–	–	–	0.385

Therefore, the matrix $\beta_{(10;,,)}$ is given by

i	$j=2$	$j=3$	$j=4$
1	0.319	0.328	0.288
2	–	1.171	0.435
3	–	–	0.471

and the vector $\alpha_{(10;.)}$ is

	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$
$\alpha_{(10;j)}$	0.589	0.311	0.563	0.237	0.269	0.385

7. Conclusion

If we implement the Chao sampling scheme, we show that the Yates-Grundy estimator has a special structure, in the sense that most of the weights of this estimator are identical. Moreover, to compute this estimator, we need only cumulative totals of the auxiliary variable.

Appendix I: Proof of Lemma 1

This result is a direct application of the following relation given by Chao (1982):

$$\pi_{(k+1;i)} = \begin{cases} \left[1 - w_k R^*_{(k;i)}\right]\pi_{(k;i)} & \text{if } i \le k \\ w_k & \text{if } i = k+1 \end{cases} \tag{15}$$

where

$$R^*_{(k;i)} = \begin{cases} T(k;i) & \text{if } \pi_{(k;i)} = 1 \\ \dfrac{1 - T_k}{n - L_k} & \text{if } \pi_{(k;i)} < 1 \end{cases} \tag{16}$$

$$T_k = \begin{cases} \displaystyle\sum_{j \in B_k} T(k;j) & \text{if } B_k \neq \emptyset \\ 0 & \text{if } B_k = \emptyset \end{cases}$$

$$T(k;i) = \frac{1 - \pi_{(k+1;i)}}{w_k}$$

$$B_k = \{i : \pi_{(k;i)} = 1, \pi_{(k+1;i)} < 1,\ i \le k\}$$

$$L_k = \#\{i : \pi_{(k;i)} = 1\}$$

By comparing (6) and (15), the lemma is proved if we show that $R_{(k;i)} = R^*_{(k;i)}$ for all $k \geqslant n$ and for all $i \le k$. We examine two cases separately, $k = n$ and $k \geq n+1$.

Case 1. If $k = n$, $\pi_{(k;i)} = 1$ for all $i \le n$. Thus

$$R^*_{(k;i)} = \frac{1 - \pi_{(n+1;i)}}{w_n}$$

i.e., $R_{(n;i)} = R^*_{(n;i)}$ for all $i \le n$

Case 2. If $k \ge n+1$, $\pi_{(k;i)} < 1$ for all $i \le k$. As set B_k is empty, $L_k = T_k = 0$. Thus $R^*_{(k;i)} = 1/n$. Now using (8), it is clear that $R_{(k;i)} = R^*_{(k;i)}$. This completes the proof.

Appendix II: Proof of Theorem 2

First $\beta_{(k+1;i,j)}$ and $\alpha_{(k+1;j)}$ can be written in the following way:

$$\beta_{(k+1;i,j)} = \begin{cases} \dfrac{\pi_{(n+1;i)}\pi_{(n+1;j)}}{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1} - 1 & \text{if } j \le n+1 = k+1 \\ p_k\left[\beta_{(k;i,j)} + 1\right] - 1 & \text{if } j \le n+1 < k+1 \end{cases} \tag{17}$$

$$\alpha_{(k+1;j)} = \begin{cases} \dfrac{n - w_k}{n-1} - 1 & \text{if } j = k+1 > n+1 \\ p_k[\alpha_{(k;j)} + 1] - 1 & \text{if } j < k+1 > n+1 \end{cases} \tag{18}$$

Indeed, using (17), we have

$$\beta_{(n+1;i,j)} = \frac{\pi_{(n+1;i)}\pi_{(n+1;j)}}{\pi_{(n+1;i)} + _{(n+1;j)} - 1} - 1 \tag{19}$$

$$\beta_{(n+2;i,j)} = p_{n+1}[1 + \beta_{(n+1;i,j)}] - 1 \tag{20}$$

$$\vdots$$

$$\beta_{(k+1;i,j)} = p_k[\beta_{k;i,j)} + 1] - 1 \tag{21}$$

Putting (19, (20) ... (21) together, we effectively obtain (10).

If we do the same with (18), we have

$$\alpha_{(j;j)} = \frac{n - w_{j-1}}{n-1} - 1 \tag{22}$$

$$\alpha_{(j+1;j)} = p_j[\alpha_{(j;j)} + 1] - 1 \tag{23}$$

$$\vdots$$

$$\alpha_{(k+1;j)} = p_k[\alpha_{(k;j)} + 1] - 1 \tag{24}$$

Putting (22), (23) .. (24) together, we effectively obtain (11).

Now, with (17) and (18), it will be easier to prove Theorem 2.

Consider two cases: $j \le n+1$ and $j > n+1$.

Case 1. Suppose that $j \le n+1$. By using $\breve{\Delta}_{(n;i,j)} = 0$ and Lemma 2, we obtain

$$\breve{\Delta}_{(n+1;i,j)} = \begin{cases} \dfrac{Q_{(n;i)}Q_{(n;j)}}{Q_{(n;i)} + Q_{(n;j)} - 1} - 1 & \text{if } j < n+1 \\ \dfrac{Q_{(n;i)}}{1 - R_{(n;i)}} - 1 & \text{if } j = n+1 \end{cases} \tag{25}$$

As

$$R_{(n;i)} = \frac{1 - \pi_{(n+1;i)}}{w_n}$$

$$Q_{(n;i)} = \pi_{(n+1;i)}$$

(25) becomes

$$\breve{\Delta}_{(n+1;i,j)} = \begin{cases} \dfrac{\pi_{(n+1;i)}\pi_{(n+1;j)}}{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1} - 1 & \text{if } j < n+1 \\ \dfrac{\pi_{(n+1;i)} w_n}{\pi_{(n+1;i)} + w_n - 1} - 1 & \text{if } j = n+1 \end{cases}$$

Comparing with (17) $\breve{\Delta}_{(n+1;i,j)} = \beta_{(n+1;i,j)}$, i.e., $\breve{\Delta}_{(k+1;i,j)} = \beta_{(k+1;i,j)}$ for $k = n$.

Now for a population size $k > n$, we can prove

$$\breve{\Delta}_{(k+1;i,j)} = \beta_{(k+1;i,j)} \tag{26}$$

via induction by supposing that

$$\breve{\Delta}_{(k;i,j)} = \beta_{(k;i,j)} \tag{27}$$

As $j \leq n+1$ and $k > n$, it is clear that $j < k+1$. Thus using (27) and Lemma 2, we obtain

$$\breve{\Delta}_{(k+1;i,j)} = \frac{Q_{(k;i)}Q_{(k;j)}}{Q_{(k;i)} + Q_{(k;j)} - 1}\left[\beta_{(k;i,j)} + 1\right] - 1 \tag{28}$$

Through (7) and (8), we have

$$Q_{(k;i)} = 1 - \frac{w_k}{n}$$

By replacing the last relation in (28) and by comparing to (17), we effectively obtain

$$\breve{\Delta}_{(k+1;i,j)} = \beta_{(k+1;i,j)}$$

Case 2. Suppose that $j > n+1$. First, we note that $\breve{\Delta}_{(k+1;i,j)}$ is only defined for $k+1 \geq j$. For $k+1 = j$, Lemma 2 gives

$$\breve{\Delta}_{(k+1;i,j)} = \frac{Q_{(k;i)}}{1 - R_{(k;i)}} - 1 \tag{29}$$

As $k \geqslant n+1$,

$$R_{(k;i)} = \frac{1}{n}$$

$$Q_{(k;i)} = 1 - \frac{w_k}{n}$$

Putting the last two expressions in (29), we have

$$\breve{\Delta}_{(k+1;i,j)} = \frac{n - w_k}{n-1} - 1$$

Thus using (18), we have $\breve{\Delta}_{(k+1;i,j)} = \alpha_{(k+1;j)}$ if $k+1 = j$. Now, we end the proof via

induction. Suppose that $\breve{\Delta}_{(k;i,j)} = \alpha_{(k;j)}$ for $k \geq j$. Using Lemma 2,

$$\breve{\Delta}_{(k+1;i,j)} = \frac{Q_{(k;i)}Q_{(k;j)}}{Q_{(k;i)} + Q_{(k;j)} - 1}[\alpha_{(k;j)} + 1] - 1 \tag{30}$$

As $k \geqslant n + 1$,

$$Q_{(k;i)} = 1 - \frac{w_k}{n}$$

By replacing the last relation in (30) and by comparing to (18), we effectively obtain

$$\breve{\Delta}_{(k+1;i,j)} = \alpha_{(k+1;j)}$$

This completes the proof.

8. References

Bethlehem, J.G. and Schuerhoff, H. (1984). Second-Order Inclusion Probabilities in Sequential Sampling without Replacement with Unequal Probabilities, Biometrika, 71, 642–644.

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

Horvitz, D.G. and Thompson, D.J. (1951). A Generalization of Sampling without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

McLeod, A.I. and Bellhouse, D.R. (1983). A Convenient Algorithm for Drawing a Simple Random Sample. Applied Statistics, 32, 182–184.

Sugden, R.A., Smith, T.M.F., and Brown, R. (1996). Chao's List Sequential Scheme for Unequal Probability Sampling. Journal of Applied Statistics, 23, 413–421.

Yates, F. and Grundy, P.M. (1953). Selection without Replacement from within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society, Series B, 1, 253–261.

Received August 1996
Revised November 1997