The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 181-189

An Optimal Scheme of Inclusion Probability Proportional to Size Sampling

Sun Woong Kim[1)]

Abstract

This paper suggest a method of inclusion probability proportional to size sampling that provides a non-negative and stable variance estimator. The sampling procedure is quite simple and flexible since a sampling design is easily obtained using mathematical programming. This scheme appears to be preferable to Nigam, Kumar and Gupta's (1984) method which uses a balanced incomplete block designs. A comparison is made with their method through an example in the literature.

Keywords : Sen-Yates-Grundy variance estimator, Balanced incomplete block designs, Mathematical programming

1. Introduction

A number of methods for sampling probability non-replacement samples have been proposed. But we may encounter the problem in deciding which is preferable over other methods with respect to the utilization of the known quantities associated with the units. Many of them are discussed by Brewer and Hanif (1983) and one of the popular sampling strategies may be the use of inclusion probability proportional to size (IPPS) selection procedure in combination with the Horvitz-Thompson (1952) estimator for the population total.

Although a variety of IPPS sampling designs have been introduced, it seems that it is not a simple matter to develop an IPPS sampling scheme having some desirable properties, described in Section 2. In addition to these properties, it is not inconceivable that the stability of Sen-Yates-Grundy (1953) variance estimator is also desired.

Nigam, Kumar and Gupta (1984) proposed an IPPS sampling scheme having all of those properties and the ideas in their selection procedure are closely related to the third of the four methods of Jessen (1969). Their method uses balanced incomplete block (BIB) designs of experimental designs, and is expected that the desirable properties are satisfied and

1) Assistant Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea
E-mail : sunwk@dongguk.edu

Sen-Yates-Grundy variance estimator remains stable, as indicated by Rao and Nigam (1992). But their method heavily emphasizes the use of the experimental design approach, which is complicated to use even for the case of a small sample size.

In this paper, we propose an IPPS sampling scheme implemented by mathematical programming in order to avoid this problem and keep the desirable properties. The selection method is simple and flexible. The usefulness of the method is demonstrated through an example in the literature.

2. The Desirable Properties of IPPS Sampling

Consider a finite population $U=\{u_1, \cdots, u_N\}$ consisting of N distinct and identifiable units. Let $s=\{u_1, \cdots, u_n\}$ be a sample of fixed size n that is drawn from U according to a given sampling design, denoted by $p(s)$, with first-order inclusion probabilities $\pi_i = p(u_i \in s)$ and second-order inclusion probabilities $\pi_{ij} = p(u_i \in s \wedge u_j \in s)$.

Let y_i and x_i denote respectively, the value of the character of interest and the known quantity for the i th unit of U. Further, let $z_i = x_i / X$, where $X = \sum_{i=1}^{N} x_i$. z_i is often called the size measure of the i th unit.

Assuming $\pi_i > 0$ for all i, the well-known Horvitz-Thompson estimator $\hat{Y}_s$ of the population total $Y = \sum_{j=1}^{N} y_j$ is defined as

$$\hat{Y}_s = \sum_{i \in s} \frac{y_i}{\pi_i} \quad . \tag{2.1}$$

The variance of $\hat{Y}_s$, say $V(\hat{Y}_s)$, may be expressed as

$$V(\hat{Y}_s) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2 \quad . \tag{2.2}$$

This expression was first derived by Sen (1953) and Yates and Grundy (1953). Also, Sen-Yates-Grundy (1953) variance estimator of $V(\hat{Y}_s)$ is given by

$$\hat{V}_{SYG}(\hat{Y}_s) = \sum_{i<j \in s} \sum \pi_{ij}^{-1} (\pi_i \pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2 \quad . \tag{2.3}$$

The desirable properties in a fixed-size IPPS sampling scheme are summarized as follows:

(*i*) $\sum_{s\in S} p(s)=1$, where S is a collection of samples of U .

(*ii*) The first-order inclusion probabilities $\pi_i=p(u_i\in s)$ are strictly proportional to x_i , $i=1,\cdots,N$.

(*iii*) The second-order inclusion probabilities satisfy $\pi_{ij}=p(u_i\in s \wedge u_j\in s)>0$ for all $i\neq j$.

(*iv*) The second-order inclusion probabilities are relatively easily calculated without undue labour.

(*v*) $\pi_{ij}-\pi_i\pi_j<0$ for all $i\neq j$ so that a non-negative Sen-Yates-Grundy variance estimator is guaranteed.

As described in the next section, the stability of Sen-Yates-Grundy variance estimator may be also regarded as one of the most important properties in an IPPS sampling scheme. Then the following criterion can be considered to measure the stability of the estimator.

(*vi*) $\min(\phi_{ij})=\min(\pi_{ij}/\pi_i\pi_j)>k$, where k is a value to be far away from zero.

3. Nigam et al.'s Sampling Scheme

Nigam, Kumar and Gupta (1984) proposed an IPPS sampling scheme using BIB designs to satisfy properties (*i*)~(*v*) and provide a stable Sen-Yates-Grundy variance estimator, resulting in k to be sufficiently away from zero with respect to property (*vi*) .

In order to explain their method, define integer values $r_i=tz_i$, where t is a positive integer divisible by n . Their method consists of the following steps:

Step 1. Choose a proper value of t and obtain the values of r_i . Construct a sample space S containing $B=t/n$ samples, each of size n , such that the i th unit in the population occurs once in each of r_i samples and the i th and j th units occur together once in each of α_{ij} samples, where α_{ij} is a parameter of the BIB design and satisfies

$$\beta_{ij}/2\leq\alpha_{ij}\leq\beta_{ij}\ , \tag{3.1}$$

where $\beta_{ij}=[r_ir_j/B]$, and $[x]$ denotes the largest integer in x .

Step 2. Select a sample s randomly from S in such a way that the probability of selecting a sample is B^{-1} .

Their method is expected that property (vi) is reasonably satisfied. Hence it seems that for the stability of Sen-Yates-Grundy variance estimator, this method is better than the third method of Jessen (1969), simply method 3, that the samples are generated by a weighted system of randomization. But the method not only demands a skillful handling of the various methods of construction of BIB designs involving considerable trial and error but also needs a judicious combination of more complicated designs for the case of $n>2$.

4. The Proposed Sampling Scheme

Now we introduce a new IPPS sampling scheme as follows. Suppose we have a sample space S that is the set of all possible samples. Let us consider a different form of the Sen-Yates-Grundy variance estimator (2.3):

$$\hat{V}_{SYG}(\hat{Y}_s) = \sum_{i=1}^{N}\sum_{j>i}^{N} \Pi_{ij}(s)\,(y_i/\pi_i - y_j/\pi_j)^2 , \tag{4.1}$$

where $\Pi_{ij}(s) = a_i(s)\,a_j(s)\,\pi_{ij}^{-1}(\pi_i\pi_j - \pi_{ij})$ and $a_i(s)=1$ if $u_i \in s$ and $a_i(s)=0$ otherwise.

Since it is desirable to obtain the smaller estimated variances over all possible samples, consider a sampling design to minimize

$$\sum_{s\in S} \hat{V}_{SYG}(\hat{Y}_s) . \tag{4.2}$$

If we consider the hypothetical situation where the weight, $\Pi_{ij}(s)$ is a constant Π , then (4.1) can be written

$$\hat{V}_{SYG}(\hat{Y}_s) = \Pi \sum_{i=1}^{N}\sum_{j>i}^{N} (y_i/\pi_i - y_j/\pi_j)^2 \tag{4.3}$$

and (4.2) have the form of

$$\sum_{s\in S} \hat{V}_{SYG}(\hat{Y}_s) = \Pi \sum_{s\in S} [\sum_{i=1}^{N}\sum_{j>i}^{N} (y_i/\pi_i - y_j/\pi_j)^2] . \tag{4.4}$$

However, $\Pi_{ij}(s)=\Pi$ is impossible for all possible samples when the size measures of the units in the population are not equal. Then we may consider a set of second-order inclusion probabilities π_{ij} such that the following is approximately minimized.

$$\sum_{s\in S}\left[\sum_{i=1}^{N}\sum_{j>i}^{N}(\Pi_{ij}(s)-\Pi)\right] \tag{4.5}$$

Noting that Π is a constant, the minimization of (4.5) reduces to minimizing

$$\frac{(N-2)!}{(n-2)!\,(N-n)!}\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij}^{-1}(\pi_i\pi_j-\pi_{ij}) \tag{4.6}$$

or

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij}^{-1}\pi_i\pi_j\ . \tag{4.7}$$

This leads to formulate a mathematical programming problem to find the optimal values of the variables $\{\pi_{ij},\ j>i=1,\cdots,N\}$ with an objective function of (4.6) or (4.7) and certain constraints to satisfy those desirable properties of IPPS sampling design, as mentioned in Section 2. Accordingly, we propose an optimal IPPS sampling design problem consisting of (4.8), (4.9) and (4.10) below in order to find a sampling design, $p(s)$, comparable to the one by Nigam, Kumar and Gupta (1984). Note that the objective function in (4.8) corresponds to (4.7), which is the simpler form than (4.6).

$$\text{Minimize}\ \sum_{i=1}^{N}\sum_{j>i}^{N}\{p(u_i\in s\wedge u_j\in s)\}^{-1}p(u_i\in s)\,p(u_j\in s) \tag{4.8}$$

subject to the following linear constraints:

$$\sum_{j\neq i}^{N}p(u_i\in s\wedge u_j\in s)=(n-1)\,p(u_i\in s)\ , \tag{4.9}$$

$$k\,p(u_i\in s)\,p(u_j\in s)\le p(u_i\in s\wedge u_j\in s)\le p(u_i\in s)\,p(u_j\in s)\ , \tag{4.10}$$

where $p(u_i\in s)=nz_i$ and $0<k<1$.

The first constraint (4.9) indicates the relation between the first-order inclusion probabilities and the second-order inclusion probabilities with respect to sampling design $p(s)$. The second constraint (4.10) ensures that the desirable properties of (ii), (iii), (v) and (vi) are exactly satisfied. Note that an appropriate value of k is the maximum in solving the design problem.

Since the first-order inclusion probability $p(u_i\in s)$ is a fixed value, we can readily find a solution to the sampling design problem above by employing any mathematical programming computer software including SAS/OR. In the result, the property (iv) is reasonably satisfied.

5. An Illustration

Consider a population of 4 units in Table 5.1 given by Yates and Grundy (1953, page 255) to show the practical utility of the proposed method. Two units are selected with the inclusion probabilities $\pi_i = p(u_i \in s)$ proportional to the size measure z_i. Note that $\pi_{ij} = p(u_i \in s \wedge u_j \in s) = p(s)$ when the sample size is 2.

Table 5.1.
Size measures of four units

Unit	z_i
1	0.1
2	0.2
3	0.3
4	0.4

Table 5.2 presents the sampling designs obtained by the proposed sampling scheme using mathematical programming approach, Jessen's (1969) method 3 and corresponding to the following three choices in Nigam, Kumar and Gupta's (1984) method.

$$t=30, r_1=3, r_2=6, r_3=9, r_4=12, B=15$$
$$t=40, r_1=4, r_2=8, r_3=12, r_4=16, B=20$$
$$t=100, r_1=10, r_2=20, r_3=30, r_4=40, B=50$$

As seen in Table 5.2, the sampling designs from Jessen's method 3 and Nigam, Kumar and Gupta's method when $t=30$ are almost same, while the others are quite different.

Table 5.2. Sampling plan according to different sampling schemes

Sample s	$p(s)$ Method				
	K	J	$N(t=30)$	$N(t=40)$	$N(t=100)$
(1, 2)	0.0365	0.0666	0.0667	0.0500	0.0400
(1, 3)	0.0545	0.0667	0.0667	0.0500	0.0600
(1, 4)	0.1090	0.0667	0.0667	0.1000	0.1000
(2, 3)	0.1090	0.0666	0.0667	0.1000	0.1000
(2, 4)	0.2545	0.2667	0.2666	0.2500	0.2600
(3, 4)	0.4365	0.4667	0.4666	0.4500	0.4400

Note. K: Proposed method, J : Jessen's (1969) method, N: Nigam et al.'s (1984) method
The sampling plans for schemes J and N are from Jessen (1969, page 183), and Nigam et al. (1984, pages 566-567), respectively.

A comparison of the values of ϕ_{ij} in the desirable property (vi) for those methods in Table 5.2 is shown in Table 5.3. Since all values of ϕ_{ij} in Table 5.3 are less than one, the property (v) indicating the non-negativity of variance estimators is guaranteed for those methods.

Table 5.3. Values of ϕ_{ij} according to different sampling designs

Sample s	ϕ_{ij}				
	Method				
	K	J	$N(t=30)$	$N(t=40)$	$N(t=100)$
(1, 2)	0.456	0.833	0.833	0.625	0.500
(1, 3)	0.454*	0.556	0.556	0.417*	0.500
(1, 4)	0.681	0.417	0.417	0.625	0.625
(2, 3)	0.454*	0.278*	0.278*	0.417*	0.417*
(2, 4)	0.795	0.833	0.833	0.781	0.813
(3, 4)	0.909	0.972	0.972	0.938	0.917

Note. * : $\min(\phi_{ij})$

See notes for Table 5.2.

It can also be seen from Table 5.3 that the proposed method would provide more stable Sen-Yates-Grundy variance estimator than in Nigam, Kumar and Gupta's method or Jessen's method 3, due to the value of $\min(\phi_{ij})$ farthest away from zero. Although the results using BIB designs in Nigam, Kumar and Gupta's method can be gradually improved by a suitable choice of t, repeating for choosing a proper value of t does not appear to be preferable in practice.

In addition, the calculations of the estimated variances may be illustrated for a population with $y_1=0.5$, $y_2=1.2$, $y_3=2.1$ and $y_4=3.2$ given by Yates and Grundy (1953, page 255). Table 5.4 presents the values of the estimated variances for individual samples. The table also shows the values of the weighted means (WM) and the coefficients of variation (CV) for those variance estimates according to different sampling designs. It is observed that the proposed method provides the smaller coefficient of variation as well as the smaller weighted mean than that provided by the other methods. Note that coefficient of variation especially indicates the level of stability of the variance estimator.

Table 5.4. Estimated variances, weighted means and coefficients of variation according to different sampling designs

Sample	$\hat{V}_{SYG}(\hat{Y}_s)$				
s	K	J	$N(t=30)$	$N(t=40)$	$N(t=100)$
(1, 2)	0.298	0.050	0.050	0.150	0.250
(1, 3)	1.202	0.799	0.799	1.400	1.000
(1, 4)	1.053	3.147	3.147	1.350	1.350
(2, 3)	0.300	0.651	0.650	0.350	0.350
(2, 4)	0.257	0.200	0.200	0.280	0.231
(3, 4)	0.025	0.007	0.007	0.017	0.023
WM	0.300	0.367	0.367	0.325	0.310
CV	1.244	2.127	2.127	1.402	1.349

Note. WM: $E[\hat{V}_{SYG}(\hat{Y}_s)]=\sum_{s\in S}\hat{V}_{SYG}(\hat{Y}_s)p(s)$

CV: $CV(\hat{V}_{SYG}(\hat{Y}_s))=\sqrt{E[\hat{V}_{SYG}(\hat{Y}_s)]^2-(E[\hat{V}_{SYG}(\hat{Y}_s)])^2}/E[\hat{V}_{SYG}(\hat{Y}_s)]$

See notes for Table 5.2.

6. Concluding Remarks

We have proposed an optimal IPPS sampling scheme using a mathematical programming. This sampling scheme is simple to use and optimal in the sense of satisfying the desirable properties of an IPPS design including the stability of variance estimator. Furthermore, for any sample sizes including the case of $n>2$, the proposed method can easily be implemented.

But the proposed method would be computer intensive since it involves the calculation of $N(N-1)/2$ values of the second-order inclusion probabilities. However, the method may be recommended to select a small number of primary sampling units from each of strata in stratified multistage sampling.

References

[1] Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities*, New York: Springer-Verlag.

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

[3] Jessen R. J. (1969). Some methods of probability non-replacement sampling, *The Journal of the American Statistical Association*, Vol. 64, 175-193.

[4] Nigam, A. K., Kumar, P. and Gupta, V. K. (1984). Some methods of inclusion probability proportional to size sampling, *Journal of the Royal Statistical Society*, Series B, Vol. 46, 564-571.

[5] Rao, J. N. K. and Nigam, A. K. (1992). Optimal controlled sampling: a unified approach, *International Statistical Review*, Vol. 60, 89-98.

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

[7] Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata and with probability proportional to size, *Journal of the Royal Statistical Society*, Series B, Vol. 15, 253-261.

[Received August 2004, Accepted February 2005]