

5-1-2009

Bias in Stabilized Sieve Sampling


Liming Guan

University of Hawaii at Manoa, lguan@hawaii.edu

John P. Wendell

University of Hawaii at Manoa, john.wendell@hawaii.edu

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Guan, Liming and Wendell, John P. (2009) "Bias in Stabilized Sieve Sampling," *Journal of Modern Applied Statistical Methods*: Vol. 8 : Iss. 1 , Article 23.

DOI: 10.22237/jmasm/1241137320

Available at: <http://digitalcommons.wayne.edu/jmasm/vol8/iss1/23>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Bias in Stabilized Sieve Sampling

Liming Guan John P. Wendell
University of Hawaii at Manoa

The stabilized sieve sample selection method (SSM) is considered to be a probability proportional to size (PPS) sampling method with an unbiased estimator (Horgan 1997, 1998). This article demonstrates that SSM does not select items with PPS and that the point estimator is biased.

Key words: Sampling with probability proportional to size; Hansen-Hurwitz estimator; Horvitz-Thompson estimator.

Introduction

Consider a situation where it is desired to make an inference about an unknown population parameter, Y , such that

$$Y = \sum_{I=1}^N y_I \quad (1)$$

where N is the population size, $I = (1, 2, \dots, N)$, and y_I is unknown but can be determined exactly by applying some procedure. An unbiased estimate of Y can be obtained when sampling with replacement using the Hansen-Hurwitz estimator (Brewer & Hanif 1983, p. 5)

$$\hat{Y}_{HH} = \frac{1}{n} \sum_i^n \frac{y_i}{p_i} \quad (2)$$

where n is the sample size, y_i is the value y_I that is determined for the i th item in the sample, p_i is the probability of inclusion as the i th item in the sample of the population item I and p_i is the value of p_I for the i th item selected for the sample. Note that under sampling with replacement an individual population item, I , can be included in the sample more than once.

Liming Guan is an Associate Professor of Accounting in the Shidler College of Business. Email: lguan@hawaii.edu. John P. Wendell is a Professor of Accounting in the Shidler College of Business. Email: john.wendell@hawaii.edu.

When sampling without replacement, an unbiased estimate of Y can be obtained using the Horvitz-Thompson estimator (Brewer & Hanif 1983, p. 6)

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (3)$$

where π_i is the probability of inclusion in the sample of the population item I and y_i is the value of y_I for the i th item in the sample.

Equations (2) and (3) are general and allow for an unbiased estimate of Y regardless of how p_i or π_i are determined. For sampling with equal probabilities $p_i = 1/N$ and $\pi_i = n/N$. Sampling with unequal probabilities is often a good choice and may be the only possible method given the sampling frame. Examples of sampling with unequal probabilities are stratified sampling and cluster sampling. Another method for sampling with unequal probabilities is probability proportional to size (PPS) sampling. The size variable can be any variable x for which every x_i satisfies

$$0 < x_i < \frac{X}{n} \text{ where } X = \sum_{I=1}^N x_I. \quad (4)$$

The right side of the inequality is a requirement only when sampling without replacement. If these conditions are met then a PPS sample can be drawn by setting

$$p_i = \frac{x_i}{X} \quad (5)$$

when sampling with replacement and

$$\pi_I = n \frac{x_I}{X} \tag{6}$$

when sampling without replacement.

PPS sampling methods are generally applicable to any population where it is desired to estimate Y using either (2) or (3) and there is a size variable available conforming to (4). This article examines the properties of two such methods, the sieve method and the stabilized sieve method (SSM).

Sieve Sampling

The sieve method is a PPS sampling without replacement method that was developed by Rietveld (1978, 1979a,b). The presentation of the method given here is based on Horgan (1998). A population item is selected for inclusion in the sample if it satisfies the inequality

$$r_I \leq x_I \tag{7}$$

where r_I is a random variable uniformly distributed on the interval $(0, X/n)$ and each r_I is independently generated. It is important to note that the realized sample size, n_r , is a random variable that will not always be the same as n . Equation (3) with the sum over n_r and π_I defined as in (6) will yield an unbiased estimate of Y . The properties of the sieve method and the SSM will be illustrated by sampling from a hypothetical population with $N = 5$ and $n = 2$ used by Wright (1991) to demonstrate that systematic PPS samples lose their PPS property when augmented by systematically sampling the remaining population. The details of this population are given in Table 1.

Table 1: Test Population. $N = 5, n = 2, X = 20$.

I	x_I	π_I	$1 - \pi_I$	p_I
1	2	0.2	0.8	0.10
2	3	0.3	0.7	0.15
3	4	0.4	0.6	0.20
4	5	0.5	0.5	0.25
5	6	0.6	0.4	0.30

Table 2: Probabilities of sample outcomes for the test population in Table 1 for sieve sampling. Column j is an identification variable for each of the 32 outcomes. The second column indicates which population items were included in a particular sample outcome and p is the probability of that outcome.

j	I_s	p
1	Null	0.0672
2	1	0.0168
3	2	0.0288
4	3	0.0448
5	4	0.0672
6	5	0.1008
7	1,2	0.0072
8	1,3	0.0112
9	1,4	0.0168
10	1,5	0.0252
11	2,3	0.0192
12	2,4	0.0288
13	2,5	0.0432
14	3,4	0.0448
15	3,5	0.0672
16	4,5	0.1008
17	1,2,3	0.0048
18	1,2,4	0.0072
19	1,2,5	0.0108
20	1,3,4	0.0112
21	1,3,5	0.0168
22	1,4,5	0.0252
23	2,3,4	0.0192
24	2,3,5	0.0288
25	2,4,5	0.0432
26	3,4,5	0.0672
27	1,2,3,4	0.0048
28	1,2,3,5	0.0072
29	1,2,4,5	0.0108
30	1,3,4,5	0.0168
31	2,3,4,5	0.0288
32	1,2,3,4,5	0.0072

BIAS IN STABILIZED SIEVE SAMPLING

These sample outcome probabilities are calculated as

$$p_j = \prod_{I \in s_j} \pi_I \prod_{I \notin s_j} 1 - \pi_I \quad (8)$$

where s_j is the j th sample outcome in Table 2. For example, the probability of getting sample outcome 11, item 2 and 3, is $0.3 \times 0.4 \times 0.8 \times 0.5 \times 0.4 = 0.0192$. That the sieve method is indeed PPS for this population can be checked by summing the probabilities for each sample outcome containing a particular population item, I , and verifying that it is equal to the value for π_I in Table 1.

Table 3 shows the probability of achieving a particular n_r . These probabilities can be calculated from Table 2 by summing all probabilities for outcomes of a given size.

Table 3: Probabilities of a realized sample size for the test population in Table 1.

n_r	p
0	0.0672
1	0.2584
2	0.3644
3	0.2344
4	0.0684
5	0.0072

Table 4 shows the probabilities of inclusion in n_r for each combination of population item and realized sample size for the test population in Table 1. These conditional probabilities are not proportional to x_I .

Table 4: Conditional probabilities of inclusion.

I	n_r					
	0	1	2	3	4	5
1	0	0.0650	0.1658	0.3242	0.5789	1
2	0	0.1115	0.2700	0.4863	0.7544	1
3	0	0.1734	0.3908	0.6314	0.8421	1
4	0	0.2601	0.5247	0.7389	0.8947	1
5	0	0.3901	0.6487	0.8191	0.9298	1

The stabilized sieve method (SSM) (Horgan 1997, 1998) is a modification of the sieve method that ensures that the final sample size is always equal to n . This section details how the method selects items for a sample and then considers the properties of point estimators of Y for samples selected using the SSM.

The SSM is selected in two stages. First an initial sample, S_1 , is selected using (7). In the second stage the sampling process is conditioned upon the number of items in S_1 (Horgan 1998)

$$S_2 = \begin{cases} S_1 + A(n - n_r) & \text{if } n_r < n \\ S_1 & \text{if } n_r = n \\ S_1 - R(n_r - n) & \text{if } n_r > n \end{cases} \quad (9)$$

where S_2 is the final sample and $A(m)$ and $R(m)$ are defined as follows: $A(m)$ selects m items one at a time (with replacement) by taking a simple random sample of one from the entire population (including items in S_1) and this item is selected for inclusion in the sample if

$$r \leq x_I \quad (10)$$

where r is a uniformly distributed random number in the interval $(0, \max(x_I)]$. The process is repeated, generating a new value for r each time, until m items are selected. All items selected using $A(m)$ satisfy (5). $R(m)$ selects m items to remove from S_1 by taking a simple random sample of size m from S_1 .

Table 5 gives the probabilities for each sample outcome and n_r for the population in Table 1 sampled using SSM. Because of the complexity of (9) some explanation of how individual cells in this table were calculated may be useful. The simplest case is when $n_r = 2$ where the values are taken directly from Table 2. Outcome 9 (2,5) with $n_r = 4$ will be used to illustrate the cases when $n_r > 2$. First, the probabilities of all the outcomes where $n_r = 4$ and both item 2 and 5 are present (outcomes 28, 29, and 31) are summed and then divided by the number of combinations of two items that can be drawn from a population of four items. This gives $(0.0072 + 0.0108 + 0.0288)/6 = 0.0078$. When $n_r < 2$, there may be more than one path to a sample outcome. For example, outcome 7 (2,3)

with $n_r = 1$ can occur when the initial sieve sample contains only item 2 or only item 3. Outcome 3 in Table 2 gives the probability of S_l containing only item 2 and Table 1 gives the value of p_l for selecting item 3 in the second stage. The probabilities that S_l contains only item 3 and that item 2 is selected in the second stage can be determined in the same manner. Thus, the probability for outcome 7 when $n_r = 1$ is $0.0288 \times 0.20 + 0.0448 \times 0.15 = 0.01248$.

Table 5 shows probabilities for sample outcomes for the test population in Table 1 for stabilized sieve sampling. Column j is an identification variable for each of the 15 outcomes. The second column indicates which population items were included in a particular sample outcome, n_r is the realized sample size in stage one and the cells contain the joint probability of the sample outcome and n_r . Table 6 provides p_l for the population in Table 1 when sampling with SSM and demonstrates that these probabilities are not PPS. These probabilities are derived from Table 5 by summing the probabilities for each sample outcome that contains a particular I divided by n , which is 2 in this case. For outcomes where I is included twice it is counted twice.

Horgan (1998, equations 8, 17, and 19) provides an estimator for Y that is conditional upon n_r :

$$\hat{Y}_s = wX \sum_{i=1}^n \frac{y_i}{x_i} \text{ where } w = \begin{cases} \frac{1}{2n - n_r} & \text{if } n_r < n \\ \frac{1}{n} & \text{if } n_r = n \\ \frac{n_r}{n^2} & \text{if } n_r > n \end{cases} \quad (11)$$

Although (11) conditions on n_r , it does not take into consideration that the probabilities of inclusion in the sample given n_r are not proportional to x_l (see Table 4). Consequently, \hat{Y}_s is a biased estimator. The expected value of \hat{Y}_s for the test population can be calculated by determining \hat{Y}_s for every cell in Table 5, multiplying the result by the probability in the cell and then taking the sum of those products. The result is $1.0917 y_1 + 1.0877 y_2 + 1.0824 y_3 + 1.0749 y_4 + 1.0637 y_5$.

Because the SSM method is a sampling with replacement method based on the sieve method, and the sieve method is a PPS method without replacement it seems reasonable to use \hat{Y}_{HH} with p_l calculated according to (5). This will also give a biased estimate of Y , the expected value of which is $0.9491 y_1 + 0.9693 y_2 + 1.0029 y_3 + 1.0111 y_4 + 0.9791 y_5$ for the Table 1 population.

It is possible to construct an unbiased estimator of Y when using the SSM with the population in Table 1. This is done by first setting each p_l to the corresponding value in Table 6 and then calculating \hat{Y}_{HH} accordingly. Unfortunately, the use of this estimator is limited to very small populations because it requires an enumeration of all 2^N possible sample outcomes for the stage 1 sieve sample.

Conclusion

The stabilized sieve method does not sample with PPS and that both \hat{Y}_s and \hat{Y}_{HH} with p_l calculated according to (5) are biased estimators of Y . Further, the calculation of the unbiased estimator is prohibitively expensive to compute for any but the smallest populations. Nonetheless, the SSM performed well in the simulations in Horgan (1997 and 1998) in comparison to the sieve method and the probability proportionate to size with replacement method (PPR).

All three of these methods have drawbacks, either the possibility of items showing up more than once in the sample (SSM, PPR) or variable sample size (sieve), or bias (SSM). Systematic PPS sampling methods utilizing a random sort of the population before application have none of these drawbacks because they select fixed size samples without replacement with probabilities that are exactly proportional to x_l (see Brewer & Hanif 1983, procedures 2 and 3). These selection methods are easily applied with modern computers if both I and x_l are available in a computer accessible file. Consequently, with these sampling frames the systematic procedures should be preferred over either the sieve, SSM, or PPR methods. However, not all sampling frames make the entire population x_l conveniently accessible by computer and the sieve, SSM, and PPR methods

BIAS IN STABILIZED SIEVE SAMPLING

may have some practical advantages with these sampling frames that offset their disadvantages. With such challenging sampling frames, the SSM method should not be ruled out simply because of the difficulty in achieving a completely unbiased estimate of Y , particularly if the population characteristics and sample sizes are similar to those used for the simulations in Horgan (1997 and 1998).

References

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities*. NY: Springer-Verlag.

Horgan, J. M. (1997). Stabilizing the sieve sample using PPS. *Auditing: A Journal of Practice and Theory*, 16, 40-51.

Horgan, J. M. (1998). Stabilized sieve sampling: A point-estimator analysis. *Journal of Business and Economic Statistics*, 16, 42-51.

Rietveld, C. (1978). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (I). *Compact: Computer en Accountant*, 15, 2-11.

Rietveld, C. (1979a). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (II) en (III). *Compact: Computer en Accountant*, 16, 2-13.

Rietveld, C. (1979b). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (IV). *Compact: Computer en Accountant*, 17, 9-18.

Wright, D. W. (1991). Augmenting a sample selected with probability proportional to size. *Auditing: A Journal of Practice and Theory*, 10, 145-158.

Table 5: Probabilities of sample outcomes for the test population in Table 1 for stabilized sieve sampling.

j	I_s	n_r					
		0	1	2	3	4	5
1	1,1	0.00067	0.00168	0	0	0	0
2	1,2	0.00202	0.00540	0.00720	0.00760	0.00380	0.00072
3	1,3	0.00269	0.00784	0.01120	0.01093	0.00480	0.00072
4	1,4	0.00336	0.01092	0.01680	0.01453	0.00540	0.00072
5	1,5	0.00403	0.01512	0.02520	0.01760	0.00580	0.00072
6	2,2	0.00151	0.00432	0	0	0	0
7	2,3	0.00403	0.01248	0.01920	0.01760	0.00680	0.00072
8	2,4	0.00504	0.01728	0.02880	0.02320	0.00740	0.00072
9	2,5	0.00605	0.02376	0.04320	0.02760	0.00780	0.00072
10	3,3	0.00269	0.00896	0	0	0	0
11	3,4	0.00672	0.02464	0.04480	0.03253	0.00840	0.00072
12	3,5	0.00806	0.03360	0.06720	0.03760	0.00880	0.00072
13	4,4	0.00420	0.01680	0	0	0	0
14	4,5	0.01008	0.04536	0.10080	0.04520	0.00940	0.00072
15	5,5	0.00605	0.03024	0	0	0	0

Table 6. Probabilities of inclusion in a sample draw for each item in the test population in Table 1 compared to the probability under PPS.

I	p_I	
	actual	PPS
1	0.09491	0.10
2	0.14540	0.15
3	0.19805	0.20
4	0.25277	0.25
5	0.30886	0.30