Inverse Simple Random Sampling with and without Replacement

Sureeporn Sungsuwan [*]

Prachoom Suwattee [**]

Abstract

In general, sampling without replacement is more precise than sampling with replacement. In this paper, we compare inverse simple random sampling with and without replacement. An unbiased estimator of proportion of the two sampling designs are considered and the precision are compared using their variance estimates. From the simulation results, the inverse sampling with replacement seems to have larger sample size especially when the population proportion is small. If the population proportion increases, the variance estimate also increases. The estimator in inverse sampling without replacement tends to have smaller variance than the one in with replacement case.

[*] Ph.D. student, Department of Statistics School of Applied Statistics National Institute of Development Administration Bangkok, Thailand 10240.

[**] Professor, Department of Statistics School of Applied Statistics National Institute of Development Administration Bangkok, Thailand 10240.

1. Introduction

In survey sampling, simple random sampling is often used because of the comfortable to design and easy to analyze (Lohr, 1999: 49). There are two procedures of selecting a sample units, sampling with replacement and sampling without replacement. Usually, sampling without replacement is preferable because all of the sample units are distinct and it leads to an estimator with a smaller variance. For sampling with replacement, the distinct units in the sample may be less than the sample size. Basu (1958) showed that in sampling with replacement, the population mean estimator from the distinct units in the sample is better than the estimator from overall units in the sample. However, Rao (1966) pointed out that if considered factors other than the efficiency, the advantage of sampling with replacement are : (1) the ease with which a sample can be drawn, (2) the simplicity of the estimator and (3) the availability of the variance estimator.

The inverse sampling is a method of sampling which requires drawings at random shall be continued until certain specified conditions dependent on the results of those drawings have been fulfilled (Kendall and Buckland, 1971 : 76). The inverse sampling are used in long history such as Haldane (1945) used inverse sampling to estimate the population proportion. Sampford (1962) proposed the inverse sampling with probability proportional to size for cluster sampling. The most applications of inverse sampling are used in study the population behaviors in case that a population units possessing the characteristics of interest are small. Raj and Khamis (1958) gave some remarks on inverse sampling with replacement that the estimating making uses only the distinct units is more efficient than uses all the units in the sample. There are many authors studied the inverse sampling with and without replacement but no comparison of the two sampling designs. So it is interesting to compare the estimator obtained from inverse sampling with and without replacement.

In this paper, the inverse simple random sampling with and without replacement are studied. We consider the unbiased estimators of the population proportion, its variance and an unbiased estimator of the variance. And finally, we use the simulation to compare the precision of the estimators.

2. Inverse Simple Random Sampling

2.1) Inverse Simple Random Sampling with Replacement

Let $U = \{u_1, u_2, ..., u_N\}$ be a finite population of known size N with the y_i-values of a study variable, $i = 1, ..., N$. In sampling procedure, each population unit has the same probability of selection, in process if the population unit has been drawn then we returned this unit to the population after its characteristics have been recorded. The same process occurs until the sample satisfies the specified conditions such as the sample contains k distinct units or the sample contains k of units with the characteristics of interest. Haldane (1945) considered the method of inverse sampling that the sampling is stopped when the k of units possessing characteristic of interest have been found in the sample. Haldane gave an unbiased estimator of the population proportion of units possessing the characteristic of interest and its variance

2.2) Inverse Simple Random Sampling without Replacement

In the case the sampling without replacement which a sampling unit is drawn and not returned to the population after its characteristics have been recorded and the sampling is continued until the k units with certain characteristics are obtained.

Salehi and Seber (2001) considered inverse sampling without replacement, they supposed that the units in the population of size N can be divided into two different groups. First group is the units that possess the characteristic of interest defined on their y -values and the second group is not. All population units have the same selection probabilities, $1/N$, in the first draw. The sampling continues without replacement until k units that possess the characteristic of interest have been selected. Salehi and Seber also gave an unbiased estimator of the population proportion and its unbiased variance estimator. The estimators are based on Murthy's estimators. Salehi and Seber (2004) applied inverse sampling to adaptive cluster sampling and gave a simple example to demonstrate the computation.

Furthermore, there are many studied in inverse sampling such as Sampford (1962) proposed the inverse sampling with probability proportional to size for cluster sampling, Pathak (1976) applied the inverse sampling to the fixed cost sampling schemes, Mukerjee and Basu (1993) applied inverse sampling for a stratified population, Christman and Lan (2001) considered inverse sampling design with and without replacement and all population units have equal probabilities of selection. In the inverse sampling design, they used stopping rule based on the number of units where their values satisfy some conditions, Greco and Naddeo (2007) considered inverse sampling design when the population units have unequal probabilities.

3. Estimation of Population Proportion

In inverse sampling with replacement, Haldane showed that if the sample size was n, then an unbiased estimator of the population proportion of units possessing the characteristic of interest was

$$\hat{p}_H = \frac{k-1}{n-1} \tag{1}$$

(Haldane,1945), where k is the number of units possessing the characteristic of interest. He also gave the variance of the estimator in (1) as

$$V(\hat{p}_H) = \frac{p^2 q}{k}\left[1 + \frac{2!q}{k+1} + \frac{3!q^2}{(k+1)(k+2)} + \ldots\right] \tag{2}$$

where p is the population proportion and $q = 1 - p$. The bounds of the variance in (2) have been given in many papers by Mikulski and Smith (1976), Sathe (1977) , Prasad (1982). Haldane gave the estimator of the variance of $\hat{p}_H$ as

$$\hat{V}_H(\hat{p}_H) = \frac{k(n-k)}{n^2(n-1)}. \tag{3}$$

This estimator is not unbiased. Finney (1949) considered the Haldane's estimator and gave an unbiased variance estimator as

$$\hat{V}_F(\hat{p}_H) = \frac{\hat{p}^2(1-\hat{p})}{k-1-\hat{p}} \tag{4}$$

In the case of sampling without replacement, Salehi and Seber (2001) presented an unbiased estimator of the population mean in inverse simple random sampling without replacement as

$$\hat{\mu} = \frac{1}{n-1}\left(\sum_{i=1}^{k}\frac{k-1}{k}y_i + \sum_{i=k+1}^{n} y_i\right), \tag{5}$$

where k is the number of unit possessing the characteristic of interest and n is the sample size. They also gave an unbiased estimator of the population proportion. This estimator is the same as Haldane's estimator.

Salehi and Seber (2001) did not give the explicit expression for the variance of $\hat{p}_{SS} = (k-1)/(n-1)$, an unbiased estimator of the population proportion but they presented an unbiased estimator of $V(\hat{p}_{SS})$ as

$$\hat{V}_{SS}(\hat{p}_{SS}) = \left(1-\frac{n-1}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-2} \tag{6}$$

The explicit expression for $V(\hat{p}_{SS})$ can be derived as a special case of Murthy's results for probability proportional to size. Murthy gave the variance of an unbiased estimator of the population total as

$$V(\hat{t}_M) = \sum_{i=1}^{N}\sum_{j<i}^{N}\left(1-\sum_{s\ni i,j}\frac{P(s\mid i)\,P(s\mid j)}{P(s)}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2 p_i p_j \tag{7}$$

(Salehi and Seber, 2001).

For inverse simple random sampling without replacement Salehi and Seber (2001) showed that

$$\frac{P(s\mid i)}{P(s)} = \frac{P(I_i=1,s)}{P(s)p_i} = \begin{cases} \dfrac{N(k-1)}{(n-1)k}, & i\in s_c \\ \dfrac{N}{n-1}, & i\in s_{\bar{c}} \end{cases} \tag{8}$$

where s_c is the set of k sample units that possess the characteristic of interest , $s_{\bar{c}}$ is the set of $n-k$ sample units that unsatisfied , s is set of all sample units where $s = s_c \cup s_{\bar{c}}$.

Let M be the number of population units that possess the characteristic of interest, the number of possible samples is $M\binom{M-1}{k-1}\binom{N-M}{n-k}(n-1)!$, so we get

$$\frac{P(s|i)P(s|j)}{P(s)} = \begin{cases} \dfrac{N^2(k-1)^2}{k(n-1)^2 M\binom{M-1}{k-1}\binom{N-M}{n-k}} & , \ i,j \in s_c \\ \dfrac{N^2(k-1)}{(n-1)^2 M\binom{M-1}{k-1}\binom{N-M}{n-k}} & , \ i \in s_c, j \in s_{\bar{c}} \ or \ i \in s_{\bar{c}}, j \in s_c \\ \dfrac{N^2 k}{(n-1)^2 M\binom{M-1}{k-1}\binom{N-M}{n-k}} & , \ i,j \in s_{\bar{c}} \end{cases} \tag{9}$$

The variance of proportion estimator can be obtained from (7) as

$$V(\hat{p}_{SS}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}^{N}\left(1-\sum_{s \ni i,j}\frac{P(s|i)P(s|j)}{P(s)}\right)\left(y_i' - y_j'\right)^2 \tag{10}$$

$$= \left(\frac{1}{N^2} - \frac{(k-1)^3(n-2)!}{M(M-1)(n-1)} - \frac{k(k-1)(n-k)(n-2)!}{M(N-M)(n-1)} - \frac{k(n-k)(n-k-1)(n-2)!}{(N-M)(N-M-1)(n-1)}\right)\sum_{i=1}^{N}\sum_{j<i}^{N}\left(y_i' - y_j'\right)^2 \tag{11}$$

where $y_i' = \begin{cases} 1 & \text{if unit i possessing the characteristic of interest} \\ 0 & \text{otherwise.} \end{cases}$

4. Comparison of the Estimators

Usually precision of the proportion estimators are compared using their variances. The precision of estimators of the population proportion in inverse simple random sampling with and without replacement can not be compare directly from the expressions of the variances. So the comparison is carried out by simulation. The simulation is based on repeated sampling from generated finite normal population of size 1,000 with mean 5 and variance 4. The population proportion of units possessing the characteristic of interest in this simulation is set to be 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6. The number *k* of units with characteristic of interest in the sample is determined from the coefficient of variation(C.V.) which is approximately equal to $\sqrt{1/(k-1)}$ (Finney, 1949). If we set C.V. equals to 50% , then *k* equals to 5 . In this simulation we have *k* equals to 5, 7, 17, 26, and 45 by setting the C.V. equals to 50%, 40%, 25%, 20% and 15%, respectively. For each situation, the 1,000 samples are drawn. For a sample *i* we calculate the proportion estimate, $\hat{p}_i$, $i = 1,...,1{,}000$ and also calculate the average of the proportion estimates,

$$\bar{\hat{p}} = \frac{1}{1{,}000}\sum_{k=1}^{1{,}000}\hat{p}_i \ . \tag{12}$$

The variance estimate is obtained from

$$v(\hat{p}) = \frac{1}{999}\sum_{i=1}^{1,000}(\hat{p}_i - \bar{\hat{p}})^2 . \tag{13}$$

The simulation results are shown in the Table below :

Averages of sample size, proportion estimates and variance estimates of unbiased estimator in inverse simple random sampling with and without replacement.

p	k	$\bar{n}$		$\bar{\hat{p}}$		$10^4\, v_{WR}(\hat{p})$	$10^4\, v_{WOR}(\hat{p})$
		WR	*WOR*	*WR*	*WOR*		
0.01	5	486.844	454.575	0.0106	0.0102	0.5258	0.4709
	7	699.055	632.863	0.0100	0.0101	0.1861	0.0843
0.05	5	98.719	98.938	0.0500	0.0495	6.5044	6.4787
	7	138.860	135.884	0.0506	0.0503	4.6651	3.8612
	17	342.566	335.466	0.0497	0.0498	1.6226	1.1132
	26	516.890	510.658	0.0503	0.0501	1.0059	0.5618
	45	901.713	883.431	0.0499	0.0500	0.5399	0.0736
0.1	5	50.143	48.570	0.0991	0.1021	24.9870	26.8308
	7	68.901	70.114	0.1011	0.0992	17.1127	15.3452
	17	169.025	167.766	0.1003	0.1004	5.5873	4.9117
	26	259.465	258.188	0.1003	0.0996	3.8978	2.5475
	45	447.240	447.063	0.1007	0.0998	2.2057	1.1996
0.2	5	25.369	25.384	0.1996	0.1976	98.8345	91.1109
	7	35.074	34.757	0.2006	0.2008	62.8200	57.4955
	17	83.985	85.248	0.2022	0.1984	20.3449	18.8063
	26	130.439	130.330	0.1994	0.1987	13.0241	11.1723
	45	225.014	224.112	0.2001	0.2001	7.6476	5.9236
0.3	5	16.538	16.336	0.3021	0.3025	162.9435	151.5511
	7	23.086	23.790	0.3027	0.2904	111.5310	90.6987
	17	56.423	56.343	0.3026	0.3009	45.3177	36.6231
	26	86.104	86.228	0.3015	0.3004	24.3352	21.8750
	45	150.155	149.705	0.2997	0.3002	14.1931	13.0479
0.40	5	12.677	12.812	0.3964	0.3903	245.5151	243.5584
	7	17.671	17.489	0.3979	0.3989	174.0450	153.3335
	17	42.431	42.213	0.4015	0.4028	63.9687	65.1906
	26	65.371	64.504	0.3974	0.4020	37.5401	34.5501
	45	112.480	112.126	0.3998	0.4010	20.9041	20.4196

(Continued)

p	k	$\bar{n}$		$\bar{\hat{p}}$		$10^4 \, v_{WR}(\hat{p})$	$10^4 \, v_{WOR}(\hat{p})$
		WR	*WOR*	*WR*	*WOR*		
0.50	5	10.109	10.009	0.4977	0.5015	312.1139	307.3230
	7	13.955	13.714	0.5026	0.5098	207.2353	204.2166
	17	34.098	33.939	0.4998	0.5003	83.5193	74.1745
	26	52.216	52.095	0.4976	0.4985	48.3275	46.8701
	45	90.333	89.595	0.4977	0.5017	26.3147	25.7364
0.60	5	8.334	8.282	0.6044	0.6030	362.1706	335.5799
	7	11.822	11.482	0.5936	0.6094	235.3978	229.8088
	17	28.474	28.362	0.5967	0.5981	87.9952	82.4596
	26	43.631	43.350	0.5963	0.6001	59.2530	59.6374
	45	75.082	74.751	0.5996	0.6014	34.2808	29.2159

Note : p is the population proportion, $\bar{n}$ is average sample size, $\bar{\hat{p}}$ is average proportion estimates, $v_{WR}(\hat{p})$, $v_{WOR}(\hat{p})$ are the variance estimates of an unbiased proportion estimators in inverse simple random sampling with and without replacement, respectively.

5. Conclusions

From the Table we see that the proportion estimates in with and without replacement are likely to be equal and they so close to the proportion p. The inverse sampling with replacement seems to have larger sample size especially when the population proportion is small. The average sample size of both sampling designs are increase if we increasing the fixed number k, which reasonable. If the population proportion increases, the small sample size are obtained. Consider the variance estimate of the estimators, we see that if the population proportion increases, the variance estimate also increases. The estimator in inverse simple random sampling without replacement tends to have smaller variance than the one in with replacement case. The conclusion on the variance estimates might depend on other factor such as the sample size.

Bibliography

Basu, D. 1958. On Sampling with and without Replacement. **Sankhya : The Indian Journal of Statistics.** 20: 287-294.

Christman, M.C. and Lan, F. 2001. Inverse Adaptive Cluster Sampling. **Biometrics**. 57: 1096-1105.

Cochran, W.G. 1977. **Sampling Techniques**. Wiley,New York.

Espejo, M.R., Singh, H.P. and Saxena, S. 2008. On Inverse Sampling with Replacement. **Statistical Papers**. 49: 133-137.

Finney, D. J. 1949. On a Method of Estimating Frequencies. **Biometrika**. 36: 233-234.

Greco, L. and Naddeo, S. 2007. Inverse Sampling with Unequal Selection Probabilities. **Communications in Statistics - Theory and Methods**. 36: 1039-1048.

Haldane, J. B. S. 1945. On a Method of Estimating Frequencies**. Biometrika**. 33: 222–225.

Kendall, M.G. and Buckland, W.R. 1971. **A Dictionary of Statistical Terms.** Hafner Publishing Company, Inc. New York.

Lohr, S.L. (1999). **Sampling : Design and analysis**. Duxbury Press.

Mikulski, P.W. and Smith, P.J. 1976. A Variance Bound for Unbiased Estimation in Inverse Sampling. **Biometrika**. 63: 216-217.

Mukerjee, R. and Basu, S.K. 1993. Inverse Sampling for Domain Estimation in a Stratified Population. **Australian Journal of Statistics**. 35(3): 293-302.

Pathak, P.K. 1976. Unbiased Estimation in Fixed Cost Sequential Sampling Schemes. **The Annals of Statistics**. 4(5): 1012-1017.

Prasad, G. 1982. Sharper Variance Upper Bound for Unbiased Estimation in Inverse Sampling. **Trabajos De Estadistica Y De Investigation Operativa**. 33: 130-132.

Raj, D. and Khamis, S.H. 1958. Some Remarks on Sampling with Replacement. **Annals of Mathematical Statistics**. 29: 550-557.

Rao, J.N.K. 1966. On the Comparison of Sampling With and Without Replacement. **Review of the International Statistical Institute**. 34: 125-138.

Salehi, M.M. and Seber, G.A.F. 2001. A New Proof of Murthy's Estimator with Applies to Sequential Sampling. **Australian and New zealand Journal of Statistics**. 43(3): 281-286.

Salehi, M.M. and Seber, G.A.F. 2004. A General Inverse Sampling Scheme and Its Application to Adaptive Cluster Sampling. **Australian and New zealand Journal of Statistics**. 46(3): 483-494.

Sampford, M.R. 1962. Methods of Cluster Sampling with and without Replacement for Clusters of Unequal Size. **Biometrika**. **49**: 27-40.

Sathe, Y.S. 1977. Sharper Variance Bounds for Unbiased Estimation in Inverse Sampling. **Biometrika**. **64**: 425-426.