PPS Network Sampling with Partial PSU Replacement

Monroe G. Sirken
National Center for Health Statistics, 3111 Toledo Rd. Hyattsville, MD 20782

Abstract
The original version of the pps network sampling with partial replacement scheme (NS1) using the original psu size measures is more efficient than single-stage conventional pps sampling with replacement scheme (CS). However, the NS1 precision improvement is small if either the psu sampling rate is small or the average psu size measure is large. Two modified pps network sampling schemes are introduced that seek to improve NS1 precision by reducing the original psu size measures. This paper compares the precision of the modified network sampling schemes relative to the NC1 and CS schemes.

Key Words: finite population corrections; psu size measures; sampling small populations.

1. Introduction

Ever since Hansen and Hurwitz (HH) (1943) introduced the theory of unequal probability sampling with replacement, numerous unequal probability schemes without replacement have been proposed to improve the precision of the HH scheme (Brewer and Hanif, 1982). However, the precision gained by pps sampling without replacement schemes is often outweighed by simplicity and practicality of pps sampling with replacement. Sanchez-Crespo and Gabeiras (SCG)(1987) introduced the theory of unequal probability sampling with partial psu replacement \and demonstrated their scheme was more efficient than pps sampling with replacement while retaining the simplicity of the HH sampling scheme.

In the pps version of the SCG scheme, for example, a sample of m of R psus is selected one at a time and every time psu i (i = 1, …,R) is selected, its size measure M_i is successively reduced by b elements where b is the largest integer $\leq MinM_i / (m-1)$ and $MinM_i$ denotes the original size measure of the smallest psu. In pps network sampling with partial psu replacement, every time a psu is selected its size measure is successively reduced by one element.

Suppose the target population contains $M = \sum_i^R M_i$ elements (j = 1, …, M) where R is the number of non overlapping psus, and M_i is the size measure of the ith psu (i = 1, …, R). A single-stage pps sample survey based on a sample of m psus estimates the total $X = \sum_i^R X_i$ where X_i is the subtotal of X contained in the ith psu. This paper compares the precision of the conventional pps sampling with replacement scheme (CS) relative to 3 pps network sampling schemes with partial psu replacement that depend on different psu size measures. Each NS scheme divides the original psu size measures by a different integer $k \leq MinM_i$ where $MinM_i$ denotes the original size of the smallest psu.

NS1 k = 1 = the original psu size measures, M_i (i= 1, …, R).

$NS\tilde{k}$ $k = \tilde{k}$ = the largest common denominator of R original psu sizemeasures. Every psu reduced size measure $M_i / \tilde{k}$ (i = 1, …,R) is an integer (i.e. is precise).

$NS\hat{k}$ $k = \hat{k}$ = the largest $k > \tilde{k}$ with the fewest reduced psu size measures, $M_i / \hat{k}$ (i = 1, …, R), that are imprecise.

2. The NS1 Scheme

M elements (j = 1,…, M) are the selection units, and a srs sample of r elements is selected without replacement which is equivalent to reducing the size measure of the psus of sampled elements by one element every time one of their elements is sampled. The NS1 estimator of X counts the quantity X_i whenever the selected element j is an element of the size measure of ith psu. The unbiased NS1 estimator of X based on a simple random sample of r elements is

$$X'_{NS1} = \frac{M}{r}\sum_j^r \frac{X_j(i)}{M_j(i)} = \frac{1}{r}\sum_j^r \frac{X_j(i)}{P_j(i)}$$

where $X_j(i) = X_i$ if j is an element of the ith psu and $X_j(i) = 0$ otherwise. Similarly, $P_j(i) = P_i = \frac{M_i}{M}$ if j is an element of the ith psu and $P_j(i) = 0$ otherwise. The unbiased CS estimator of X based on a pps sample of r psus is $X'_{CS} = \frac{1}{r}\sum_i^r \frac{X_i}{P_i}$.

The NS1 sample variance based on a sample of r elements is

$$V(X'_{NS1}) = F_{NS1} \times V(X'_{NS1})$$

where

$F_{NS1} = \frac{M-r}{M-1}$ = the NS1 finite population correction

$$V(X'_{NS1}) = V(X'_{CS}) = \frac{1}{m}\sigma^2_{CS}$$

= the CS sample variance based on a sample of r psus

and

$$\sigma^2_{CS} = \sigma^2_{NS1} = \sum_i^R \left(\frac{X_i}{P_i} - X\right)^2 P_i.$$

The NS1 design effect relative to CS is

$$Deft_{NS1} = \frac{V(X'_{NC1})}{V(X'_{CS})} = F_{NS1} = \frac{M-r}{M-1} = \frac{M}{M-1}\left(1 - \frac{r}{R} \times \frac{1}{\overline{M}}\right) < 1$$

where r/R is the psu sampling rate and $\overline{M} = M / R$ = the average psu size measure.

3. The $NS\tilde{k}$ Scheme

Let $\mu_i(\tilde{k}) = M_i / \tilde{k}$ denote the precise size measure of the ith psu and $\mu(\tilde{k}) = \sum_i^R \mu_i(\tilde{k}) = M / \tilde{k}$ = the reduced size measures summed over R psus [$\tilde{j}$ = 1, …, $\mu(\tilde{k})$]. A srs sample of r elements is selected from $M / \tilde{k}$ elements without replacement.

The $NS\tilde{k}$ unbiased estimator of X based on a sample of r elements is

$$X'_{NS\tilde{k}} = \frac{1}{r}\sum_{\tilde{j}} \frac{X_{\tilde{j}}(i)}{P_{\tilde{j}}(i)}$$

where $X_{\tilde{j}}(i) = X_i$ if $\tilde{j}$ is and element of the ith psu and $X_{\tilde{j}}(i) = 0$ otherwise. Similarly, $P_{\tilde{l}}(i) = \mu_i(\tilde{k}) / \mu(\tilde{k}) = M_i / M = P_{\tilde{l}}$ if $\tilde{j}$ is an element of the ith psu and $P_{\tilde{j}}(i) = 0$ otherwise.

The $NS\tilde{k}$ sample variance based on a sample of r elements selected without replacement is

$$V(X'_{NS\tilde{k}}) = F_{NS\tilde{k}} \times V(X'_{NS\tilde{k}})$$

where

$$V(X'_{NS\tilde{k}}) = V(X'_{CS})$$

and

$$Deft_{NS\tilde{k}} = F_{NS\tilde{k}} = \frac{\mu(\tilde{k})-r}{\mu(\tilde{k})-1} = \frac{M-\tilde{k}r}{M-\tilde{k}}$$

is the $NS\tilde{k}$ design effect relative to NC. Clearly, $F_{NS\tilde{k}} \le F_{NS1}$ and $F_{NS\tilde{k}}(\tilde{k} = 1) = F_{NS1}$.

4. The $NS\hat{k}$ Scheme

Let $\mu_i(\hat{k})$ denote the $NS\hat{k}$ size measure of these ith psu

$$\mu_i(\hat{k}) = \left\langle \frac{M_i}{\hat{k}} \right\rangle = \frac{M_i}{\hat{k}} + \varepsilon_i(\hat{k}) \quad \text{(i = 1, …,R)}$$

where $\langle y \rangle$ denotes y rounded to the nearest integer and $-\frac{1}{2} \le \varepsilon_i(\hat{k}) \le +\frac{1}{2}$ are lower and upper bounds of the rounding error, $\varepsilon_i(\hat{k}) = \mu_i(\hat{k}) - \frac{M_i}{\hat{k}}$. The ith psu size measure, $\mu_i(\hat{k})$, is precise if $\varepsilon_i(\hat{k}) = 0$. Two or more $MS\tilde{k}$ psu size measures are always imprecise. Let

$\mu(\hat{k}) = \sum_i \mu_i(\hat{k}) = \frac{M}{\hat{k}} + \varepsilon(\hat{k})$ = the reduced number of $NS\hat{k}$ elements

$[\hat{j} = 1, ..., \mu(\hat{k})]$ where $\varepsilon(\hat{k}) = \sum_i \varepsilon_i(\hat{k})$

$\pi_i = \frac{\mu_i(\hat{k})}{\mu(\hat{k})}$ = the $NS\hat{k}$ selection probability of the ith psu (i = 1, …, R), and

$\pi_i = P_i = \frac{M_i}{M}$ if and only if $\varepsilon_i = 0$.

The $NS\hat{k}$ unbiased estimator of X based on a srs of r elements is

$$X'_{NS\hat{k}} = \frac{1}{r} \sum_{\hat{j}} \frac{X_j(i)}{\pi_j(i)}$$

where $X_{\hat{j}}(i) = X_i$ if $\hat{j}$ is an element of the ith psu, and $X_{\hat{j}}(i) = 0$ otherwise. Also, $\pi_{\hat{j}}(i) = \pi_i$ if $\hat{j}$ is an element of the ith psu, and $\pi_{\hat{j}}(i) = 0$ otherwise.

The $NS\hat{k}$ sample variance is

$$V(X'_{NS\hat{k}}) = F_{NS\hat{k}} \times \frac{1}{r} \sigma^2_{NS\hat{k}}$$

where

$$F_{NS\hat{k}} = \frac{\mu(\hat{k}) - r}{\mu(\hat{k}) - 1}$$

and

$$\sigma^2_{NS\hat{k}} = \sum_i \left(\frac{X_i}{\pi_i} - X \right)^2 \pi_i \neq \sigma^2_{CS}.$$

To improve transparency, assume $\varepsilon(\hat{k}) = \varepsilon = 0$ and it follows that $\mu(\hat{k} \mid \varepsilon = 0) = M / \hat{k}$, $\pi_i(\hat{k} \mid \varepsilon = 0) = P_i + \frac{\hat{k}}{M} \varepsilon_i(\hat{k})$. Making these substitutions

$$F_{NS\hat{k}}(\varepsilon = 0) = \frac{M - r\hat{k}}{M - \hat{k}} < F_{NS\tilde{k}}$$

and

$$\frac{\sigma^2_{NS\hat{k}}(\varepsilon=0)}{\sigma^2_{CS}} = \frac{\sum_i^R \left(\Delta_i + \frac{\hat{k}}{M}\varepsilon_i(\hat{k})\right)^2 / \pi_i}{\sum_i^R \Delta_i^2 / P_i}$$

where $\Delta_i = P_i - \frac{X_i}{X}$ and $\sum_i^R \Delta_i = 1$. Thus,

$$Deft_{NS\hat{k}}(\varepsilon=0) = F_{NS\hat{k}}(\varepsilon=0) \times \frac{\sigma^2_{NS\hat{k}}(\varepsilon=0)}{\sigma^2_{CS}}$$

where the finite population factor is always < 1, and the ratio of the population variances is more or less than 1 respectively depending on whether the covariance of Δ_i and $\frac{\hat{k}}{M}\varepsilon_i(\hat{k})$ (i = 1, …,R) is positive or negative. Hence, it is infeasible to predict $Deft_{NS\hat{k}}(\varepsilon=0)$ because Δ_i (i =1, …, R) are unknown.

5. Discussion

The NS1 design effect relative to CS is

$$Deft_{NS1} = \frac{V(X'_{NC1})}{V(X'_{CS})} = \frac{M-r}{M-1} = \frac{M}{M-1}\left(1 - \frac{r}{R} \times \frac{1}{\overline{M}}\right) < 1$$

where r / R is the psu sampling rate and $\overline{M} = M / R$ = the average psu size measure. The NS1 is always more efficient than the CS but the gain in precision is very small if $\overline{M}$ is large and m / R, the psu sampling rate, is small. Thus, in most pps sampling applications the efficiency gains will be small. The gains might be more substantial, however, in pps sampling of small or rare populations where m/R could be relatively large if the populations are partitioned into relatively few psus and $\overline{M}$ could be relatively small if the psu size measures are commensurate with small population sizes.

This paper introduces 2 modified pps network sampling schemes, $NS\tilde{k}$ and $NS\hat{k}$, that seek to improve the NS1 design effects relative to NC by reducing $\overline{M}$ and they reduce $\overline{M}$ by dividing the every original psus size measure M_i (i = 1, …,R) by positive integers $\tilde{k}$ and $\hat{k}$ respectively.

The $NS\tilde{k}$ design effects relative to CS is

$$Deft_{NS\tilde{k}} = \frac{V(X'_{NS\tilde{k}})}{V(X'_{CS})} = \frac{M}{M-1}\left(1 - \frac{r}{R} \times \frac{\tilde{k}}{\overline{M}}\right) \le Deft_{NS1}$$

where the integer $\tilde{k} \le MinM_i$ is the largest common denominator of the R original psu size measures and $MinM_i \le \overline{M}$ is the original size measure of the smallest psu. The smallest and largest design effects are respectively $Deft_{NS\tilde{k}}(\tilde{k}=1) = Deft_{NS1}$ and

$Deft_{NS\tilde{k}}(\tilde{k} = \overline{M}) = \dfrac{R-m}{R-1} = F_{CS\overline{w}}$ = the finite population correction in equal probability psu sampling without replacement. This largest design effect suggests that the $NS\tilde{k}$ scheme could possibly serve as a compromise between pps sampling with replacement and equal probability sampling without replacement when $P_i \approx \dfrac{1}{R}$ (i = 1, …,R). Stratification is another $NS\tilde{k}$ option worthy of consideration whenever $\hat{k} / \overline{M}$ is small.

In $NS\hat{k}$, $\hat{k} \le MinM_i$ the largest integer > $\tilde{k}$ that has the fewest imprecise reduced psu sizes. $NS\hat{k}$ has the potential to be more efficient than $NS\tilde{k}$ if the $Cov(\Delta_i, \varepsilon_i) > 0$ where $\Delta_i = \frac{M}{M} - \frac{X_i}{X}$ and $\varepsilon_i = \left\langle \frac{M_i}{\hat{k}} \right\rangle - \frac{M_i}{M}$ and $\langle y \rangle$ denotes y rounded to the nearest integer. However, the potential gains of the $NS\hat{k}$ scheme will probably not be realized until suitable models are developed to estimate $Cov(\Delta i, \varepsilon_i)$ in the absence of information about the Δ_i (i = 1, …, R),.

Acknowledgements

I want to thank Iris Shimizu for her invaluable assistance in formatting this paper for publication.

References

Brewer and Hanif (1982). *Sampling with Unequal Probabilities.* New York, Springer-Verlag.

Hansen, M.H. and Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. *Annals of Math. Stat.* 14, 333-362.

Sanchez-Crespo, J.L. (1997). A Sampling Scheme with Partial Replacement. *Journal of Official Statistics*, 4, 327 -339.

Sirken, Monroe (2001). The Hansen-Hurwitz Estimator Revisited: PPS Sampling without Replacement. *ASA Proceedings of the Survey Methods Section*.