RELATIVE EFFICIENCY OF TWO TWO-STAGE SAMPLE DESIGNS

Mohammad A. Chaudhary
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7400

KEY WORDS: Multistage Cluster Sampling, Unequal Probability, Variance Estimation

Cluster sampling procedures with unequal probabilities and without replacement described by Brewer and Hanif (1983) offer survey statisticians increased efficiency and administrative convenience when compared to procedures with equal probabilities and with replacement. Unbiased estimation of variance in cluster sampling frequently involves the computation of marginal and joint inclusion probabilities of units and within first stage components of variance. Historically, these computations have been highly intensive for computer implementation. However recent advances in high speed data processing have made possible the computation of valid estimates of sampling variances on a regular basis even for highly complex designs. This paper provides an empirical study of the relative efficiency of two two-stage sample designs. Design 1 uses Sampford (1967) procedure at the first stage and pps-systematic at the second, where as Design 2 pps-systematic sampling at both the stages.

1. Introduction

The probability proportional to size sampling has been extensively used in the sample designs of large demographic, public health, and agricultural surveys in the last few decades. Apart from increased efficiency, probability proportional to size (PPS) sampling provides a good control over the variability in sample size as well as administrative convenience in data collection by equal distribution of workloads in each psu. Further it is well known that selection with equal probability (SRS) and with replacement leads to less precise estimators than those based on sampling without replacement, the proportional reduction in variance is given by n/N. It is therefore intuitive to expect that a similar advantage in precision can be achieved when sampling without replacement is applied in conjunction with unequal probability sampling. A large number of sampling procedures with unequal probability and without replacement are available in the literature. Brewer and Hanif (1983) provided an excellent review of 50 such procedures. Bayless and Rao (1970) and others have compared the relative performance several of these procedures analytically as well as empirically. Previous empirical investigations used very small populations and were restricted to single stage sampling.

This paper provides an empirical study of the efficiency of two well known unequal probability without replacement sampling schemes in two stage setting and using a much bigger population. A subset of the National Health Interview Survey (NHIS) 1991 person data on the recoded age and family size variables is used to create a dataset to serve as the population. Design I uses Sampford (1967) scheme at the first stage and replicated pps-systematic at the second stage and Design 2 uses pps-systematic at the first stage and replicated pps-systematic at the second. For Design 2, the approximation of the variance estimator when pps-systematic sampling due to Hartley and Rao (1962) is used to estimate the variance at stage 1. Since the Yates-Grundy variance estimator is not appropriate to use with pps-systematic method, replicated pps-systematic sampling is used at the second stage for both the designs.

The pps-systematic and Sampford's procedures are briefly described in section 2. Section 3 deals with the estimation of variance for the two sample designs. A brief description of the data and the results are provided in section 4. The findings of the study are summarized in section 5. For the sake of simplicity all the results presented here would refer to a single stratum. It is also assumed that y_i are only subject to sampling errors.

2. Sampling Schemes

2.1 PPS-Systematic Procedure

Given a finite population of N units with characteristics y_t, $(t = 1, 2 .. N)$, whose total $Y = y_1 + y_2 + .. + y_N$ is to be estimated. The N units are arranged in some order preferably incorporating some desired implicit stratification. Let $p_t = x_t / \sum_{t=1}^{N} x_t$ where x_t denotes the measure of size for the t^{th} unit in the population. The cumulative totals of np_j, $T_j = \sum_{i=1}^{j} np_i$, $T_0 = 0$, are obtained. A variate d is randomly selected as a starting point with $0 \le d < 1$ and then n units are selected whose indices j satisfy $T_{j-1} \le d + k \le T_j$ for $k = 0, 1, 2, ..., n-1$.

PPS systematic is one of the most widely used methods of sampling with unequal probabilities and without replacement. It is easy to implement and is available for any sample size n. If properly used, it may result in very efficient design because of utilizing any implied or hidden stratification in the population. The inclusion probability of unit i is $\pi_i = np_i$. For this scheme, the Horvitz-Thompson estimator

$$\hat{Y} = \sum\nolimits_{i=1}^{n} y_i / \pi_i \tag{2.1}$$

is unbiased for population total Y. The joint inclusion probabilities π_{ij} will be zero for certain pairs of observations, thereby defeating the unbiasedness property of the Sen-Yates-Grundy variance estimator. Several alternative biased but simple as well as unbiased estimators of variance have been proposed in the literature. See Iachin (1982) and Murthy and Rao (1988). The estimator of variance using replicated systematic samples is unbiased, easy to use, and if efficiently employed, may produce more precise estimates than by other methods requiring the same amount of labor (Tornqvist, 1963). This method is used at the second stage of both the sample designs.

Since a small number of psu's are usually selected, replicated sampling is not appropriate to use at stage 1. Hartley and Rao (1962) provided the following approximation which is correct to O(N) using the assumptions that the units are randomly arranged and $np_i < 1$.

$$V(\hat{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j<i}^{n}\left(1-(\pi_i+\pi_j)+\sum_{i=1}^{N}\frac{\pi_i^2}{n}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \tag{2.2}$$

2.2 Sampford's Rejective Procedure

First unit is selected with probability p_i and all subsequent units are drawn with probabilities proportional to $\lambda_i = p_i/(1-np_i)$ with replacement. If any unit is selected twice the whole sample is rejected and the process is repeated until a sample of distinct units is obtained. The procedure is not as tedious as it appears because a sample is discarded as soon as a duplicate unit appears.

Let $S(m)$ denote a set of $m \leq N$ different units i_1, i_2, .. i_m and define

$$L_m = \sum_{S(m)} \lambda_{i1}\lambda_{i2}..\lambda_{im},\ 1 \leq m \leq N \text{ and } L_0 = 1, \tag{2.3}$$

where summation is over all possible sets of m distinct units drawn from the population. Also $L_m(\bar{i})$ and $L_m(\bar{i},\bar{j})$ are defined in the same way as L_m, except the units i and units i and j respectively are excluded from the population. Under this scheme the probability of selecting a particular sample consisting of units i_1, i_2, .. i_n is specified as

$$P\{S(n)\} = nK_n\lambda_{i1}\lambda_{i2}\ldots\lambda_{in}\left(1-\sum_{u=1}^{n}p_{iu}\right), \tag{2.4}$$

$$\text{where} \quad K_n = \left(\sum_{t=1}^{n}\frac{tL_{n-t}}{n^t}\right)^{-1} \tag{2.5}$$

For this sampling procedure $\pi_i = np_i$ and

$$\pi_{ij} = K_n\lambda_i\lambda_j\sum_{t=2}^{n}\frac{\left[t-n(p_i+p_j)\right]L_{n-t}(\bar{i},\bar{j})}{n^{t-2}} \tag{2.6}$$

The formula for π_{ij} is complex for $n > 2$ but can be computed on a routine basis given the availability of modern computing facilities. This procedure is simple to operate. The Yates-Grundy variance estimator is unbiased.

3. Unbiased Variance Estimation in Multistage Survey

Rao (1975) approach for multistage variance estimation is adopted for this study. A general linear unbiased estimator of population total Y may be written as

$$\hat{Y} = \sum\nolimits_{i=1}^{n} a_{is}\hat{Y}_i, \tag{3.1}$$

where a_{is} are real numbers predetermined for each sample s. Assuming that an unbiased estimator Y for single stage cluster sampling is available in the form

$$v(\hat{Y}_I) = \sum_{i=1}^{n} b_{is}Y_i^2 + \sum_{i<j}^{n} c_{ijs}Y_iY_j, \tag{3.2}$$

where b_{is} and c_{ijs} are real numbers predetermined for every sample s. $\hat{Y}_I$ is the etimator of Y for single stage sampling. It is further assumed that an unbiased estimator $\hat{\sigma}_{is}^2$ (based on sampling at the second and subsequent stages) of σ_{is}^2 is available. If σ_{is}^2 is assumed to be fixed and independent of s i.e., $\sigma_{is}^2 = \sigma_i^2$ for every s, then the unbiased estimator of $V(\hat{Y})$ is

$$v_A(\hat{Y}) = v(\hat{Y}_I) + \sum_{i=1}^{n} a_{is}\hat{\sigma}_i^2 \tag{3.3}$$

However if σ_{is}^2 is considered to be a random variable then the unbiased estimator of $(\hat{Y})$ is given as

$$v_B(\hat{Y}) = v(\hat{Y}_I) + \sum_{i=1}^{n} (a_{is}^2 - b_{is}) \hat{\sigma}_{is}^2 \qquad (3.4)$$

This formula is generally applicable irrespective of the nature of σ_{is}^2 and will be used for the estimation of variance in this study.

A two stage sample is drawn as follows. At the first stage, n psu's are selected from a total of N using one of the UEPWOR sampling schemes and then a sample of m_i ssu's are chosen from a total of M_i from the i^{th} selected psu by replicated pps-systematic sampling in the form of r_i systematic sub-samples of ssu's with sampling interval $r_i k_i$ such that $M_i = m_i k_i$. The m_i is fixed for this study, however it may be chosen to make the estimator self-weighting. For the estimator (3.1),

$$\hat{Y}_i = \sum_{j=1}^{r_i} \hat{Y}_{ij} / r_i \text{, and } \hat{Y}_{ij} = \sum_{l=1}^{[m_i/r_i]} y_{ijl} / [m_i / r_i] p_{ijl} \text{,}$$

where y_{ijl} is the total for the l^{th} ssu, selected in the j_{th} replicate from the i^{th} sample psu and p_{ijl} is the proportion of the size of the particular unit within the i^{th} psu. Using (3.4) the unbiased estimator of $(\hat{Y})$ is given by

$$v(\hat{Y}) = v(\hat{Y}_c) + \sum_{i=1}^{n} \left(a_{is}^2 - b_{is}\right) \frac{k_i - 1}{k_i r_i (r_i - 1)} \sum_{j=1}^{r_i} \left(\hat{y}_{ij} - \hat{Y}_i\right)^2 \qquad (3.6)$$

where $v(\hat{Y}_c)$ denotes the copy of $v(\hat{Y}_I)$ obtained by replacing y_i's by $\hat{Y}_i$. We need to extract the coefficients a_{is} and b_{is} for the variance estimators used at the first stage.

The Design 1 uses Sampford's procedure at stage 1 and pps-systematic at stage 2. $\pi_i = np_i$, so that $a_{is} = 1/np_i$. Since $v(\hat{Y}_c)$ is the Yates-Grundy variance estimator and from (3.2),

$$b_{is} = \frac{1}{\pi_i^2} \sum_{j \neq i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \qquad (3.6)$$

where π_{ij} is given by (2.6).

The Design 2 used pps-systematic sampling at stage 1 and replicated pps-systematic at stage 2. $\pi_i = np_i$, so that $a_{is} = 1/np_i$.

From $v(\hat{Y}_c)$ as given by (2.2)

$$b_{is} = \frac{1}{(n-1)\pi_i^{\,2}} \sum_{j \neq i}^{n} \left(1 - (\pi_i + \pi_j) + \sum_{t=1}^{N} \frac{\pi_t^2}{n} \right) \qquad (3.8)$$

4 The Data and Results

For the empirical evaluation of the efficiency of the two sample designs, data on age and family size of the 26638 respondents from 69 MSA non-self-representing psu's in National Health Interview Survey (1991)[1] were used. A total of 30 psu's were created by combining NHIS sample psu's to get adequate number of ssu's within each psu. The ssu refers to a concatenation of the processing quarter, random recode of psu number, week-census code and segment number. The modified psu's contained 42-100 ssu's each. The data were sorted first by the psu number and then by the ssu total of age or family size variable. The size measure for the selection at stage 1 is the number of respondents per psu and for stage 2 is the number of respondents per ssu. The intra-cluster correlation coefficient (rho) for age is 0.0089 and 0.0216 for family size. The population totals Y for the age and family size variables are 904293 and 88509 respectively. At stage 1, n = 5, 10 psu's are selected and at stage 2, $m_i = m = 12$ ssu's were chosen from each sample psu in four replicates of three units each. Using each of the two sample designs, 4 sets of 1000 samples were obtained and estimator of total $\hat{Y}$ and $v(\hat{Y})$ computed. The four sets correspond to four combinations of two variables and the two stage 1 sample sizes. The results are summarized in Table 1. Because the age and family size variables are only available in the recoded form, the estimates of total have no direct referent and are shown here for the purpose of comparison.

5. Conclusion

The variance figures of $\hat{Y}$ and $v(\hat{Y})$ reflect the stability of these estimators where as the means of $v(\hat{Y})$ provide a measure of the efficiency of the sample designs. For these data, Design 2 seems to perform consistently better Design 1 in terms of the stability of the estimator $\hat{Y}$. For age data, $v(\hat{Y})$ is more stable for Design 2 but there is no difference for the family size data. Design 2 is more efficient for n = 10 but almost the same for n = 5. Design 2 is also simpler to execute

[1] National Center for Health Statistics, National Health Interview Survey, 1991.

Table 1: The means and variances $\hat{Y}$ and $v(\hat{Y})$ based on 1000 samples selected with sample designs 1 & 2.

Variable	Sample Size (n)	Design 1				Design 2			
		$\hat{Y}$		$v(\hat{Y})$		$\hat{Y}$		$v(\hat{Y})$	
		Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
Age	5	908491	18.91E8	18.60E8	1.61E18	904934	14.77E8 (78.1)*	18.51E8 (99.5)	1.49E18 (92.5)
	10	905023	7.67E8	8.64E8	1.22E17	904858	6.02E8 (78.5)	8.41E8 (97.3)	9.93E16 (81.4)
Family Size	5	88466	16.89E6	15.12E6	1.04E14	88561	11.71E6 (69.3)	15.20E6 (100.5)	1.03E14 (99.0)
	10	88613	7.00E6	6.61E6	6.39E12	88469	5.71E6 (81.6)	6.64E6 (96.8)	6.39E12 (100.0)

* The relative variance for Design 2 with reference to Design 1 in percentage.

Acknowledgments: I would like to thank Drs. P.K. Sen, W.D. Kalsbeek and C.M. Suchindran of the Department of Biostatistics, University of North Carolina at Chapel Hill, for their encouragement and review of the manuscript.

6. References

Bayless, D.L., and Rao J.N.K. (1970), "An Empirical Study of Estimators and Variance Estimators in Unequal Probability Sampling (n=3 or 4)", *Journal of American Statistical Association*, 65, 1645-1667.

Brewer and Hanif, (1983), "Unequal Probability Without Replacement Sampling" Springer Verlag New York Inc.

Iachin, R. (1982), "Systematic Sampling: A Critical Review", International Statistical Review, 50, 293-303.

Murthy, M.N. (1988), "Systematic Sampling with Illustrative Examples", Handbook of Statistics 6, 147-85.

Tornqvist, L., (1963), "The Theory of Replicated Systematic Cluster Sampling with Random Start", Review of the International Statistical Institute, 31, 11-23.

Rao, J.N.K. (1975), "Unbiased Variance Estimation for Multi-stage Designs", *Sankhia, Ser. C*, **37**, 133-139.

Hartley, H.O., and Rao, J.N.K. (1962), "Sampling With Unequal Probabilities and Without Replacement", *Annals of Mathematical Statistics*, 33, 350-374.

Sampford, M.R., (1967), "On Sampling Without Replacement With Unequal Probabilities of Selection", *Biometrika*, 54, 499-513.