SAMPLING VARIANCE OF A MULTI-PHASE STRATIFIED DESIGN ON PARTIALLY EXCHANGEABLE SEQUENCES

Chun-Kai Huang [1]
University of Cape Town & University of KwaZulu-Natal
e-mail: *chun-kai.huang@uct.ac.za*
and
Delia North
University of KwaZulu-Natal
and
Temesgen Zewotir
University of KwaZulu-Natal

Key words: Multi-phase sampling, Stratification, Exchangeability, Auxiliary variable, Ratio.

Summary: In estimating the population mean of a study variable y, we can often use a ratio-type estimator when a related auxiliary variable x, with improved accessibility, is available. In cases where x is qualitative, or may be categorised, and a double sampling plan is used, we may consider a two-phase stratified sampling design. Traditionally, it is assumed that the N variables representing the readings on y are IID within and across strata. In this paper, we relax this assumption to a judgment of exchangeable sequences within each stratum, while still maintaining the assumption of independence across strata. This caters for the existence of dependence structures for within-stratum readings. We propose a methodology for estimating the variance of the ratio estimator under this scenario. Through an example, we show that this method provides a significantly more conservative estimate for the sampling variance, as compared to the standard approach.

1. Introduction

When considering the task of estimating the population mean of a study variable y, it is often the case that information on an auxiliary variable x is readily available for all units in the population. In such situations, it is common to utilise a ratio- or regression- type estimator to improve the efficiency in estimation (Cochran, 1977). However, when x is not known over the whole population, but still easier to obtain than y, we may implement a two-phase, or double, sampling design. The value of x is observed for a large sample in phase 1 and y is subsequently recorded for a subsample in phase 2. This can be generalised to cater for multiple auxiliary variables with varying levels of accessibility and correlation, where several chain-type estimators are proposed (Mukerjee, Rao and Vijayan, 1987; Singh, Singh and Shukla, 1994; Ahmed, 1998; Bhushan, Pandey and Katara, 2008; Hamad, Hanif and Haider, 2013).

[1] Corresponding author.

As a way to measure how good a sampling estimator is, the estimator variance, or mean square error in the case of biased estimators, needs to be estimated. These are usually approximated by their corresponding asymptotic expressions, which commonly assumes IID observations. A way to relax the IID condition is to take on the Bayesian approach to finite population sampling, which assumes that the observations are *exchangeable* (Ericson, 1969; Treder and Sedransk, 1996). However, this approach also requires formalisation of prior information and known sampling distributions (or at least estimates of them).

In this paper, we consider the case where x is a stratification variable, which is more easily accessible, and observations for y are obtained through phase 2 sampling from each stratum. We further assume the judgment of exchangeability within each stratum, while strata are mutually independent. This corresponds to finite population sampling without replacement. We propose a way to approximate the estimator variance under this scenario, using stationary bootstrapping at different levels of the sampling process. An example is considered which shows the standard procedure estimate underestimating the estimator variance, while our method provides an improvement.

2. Multi-phase stratified sampling

Let $U = \{1,2,\ldots,N\}$ be the index set of a finite population of size N and y be the primary variable of interest. Suppose x is an auxiliary variable related to y, which is less expensive or is easier to measure. In this situation, it is common to consider a two-phase sampling design. In the first phase a large sample $S' \subset U$ of size n' is drawn using SRSWOR and the auxiliary variable x is observed. Subsequently, a subsample $S \subset S'$ of size n is drawn, using SRSWOR, to observe y. One way of incorporating the auxiliary information into the estimation of the population mean $\bar{y}_U$, is to use a ratio estimator

$$\bar{t}_{rat} = \frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'} \,,$$

where $\bar{y}_n = n^{-1}\sum_{i\in S} y_i$, $\bar{x}_n = n^{-1}\sum_{i\in S} x_i$ and $\bar{x}_{n'} = (n')^{-1}\sum_{i\in S'} x_i$.

Often members of U can be cross-classified into groups based on the auxiliary variable; either the variable is qualitative in nature (e.g. gender), or may be categorised (e.g. age). This scenario is classically associated with stratified sampling design, with unknown population stratum sizes. Suppose that the stratification variable $x \in \{1,\ldots,H\}$ is only observed after phase 1 and samples S_h (of sizes m_h) are subsequently drawn from each stratum using SRSWOR . This results in an estimator for $\bar{y}_U$ as

$$\bar{t}_{str} = \frac{1}{n'}\sum_{h=1}^{H} n_h \bar{y}_h \,,$$

where n_h is the number of units in S with $x = h$ and $\bar{y}_h = m_h^{-1}\sum_{i\in S_h} y_i$. The variance for this estimator is given by

$$V(\bar{t}_{str}) = \left(1 - \frac{n'}{N}\right)\frac{S_y^2}{n'} + E\left[\sum_{h=1}^{H}\left(\frac{n_h}{n'}\right)^2\left(1 - \frac{m_h}{n_h}\right)\frac{s_h^{2'}}{m_h}\right] ,$$

where S_y^2 is the population variance of y and $s_h^{2'}$ is the sample variance of y in stratum h, from phase

1 if we observe them all. This can be estimated by

$$\hat{V}(\bar{t}_{str}) = \frac{N-1}{N}\sum_{h=1}^{H}\left(\frac{n_h-1}{n'-1}-\frac{m_h-1}{N-1}\right)\frac{n_h}{n'}\frac{s_h^2}{m_h}+\frac{1}{n'-1}\left(1-\frac{n'}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n'}\left(\bar{y}_h-\bar{t}_{str}\right)^2 ,$$

where s_h^2 is the sample variance of y in stratum h from phase 2 (Rao, 1973).

3. Exchangeable sequences

Under the model-based approach to sampling, the variance and estimated variance of $\bar{t}_{str}$ are derived based on the underlying assumption of IID of the random sequence $Y_1,\ldots,Y_N$ (for which $y_1,\ldots,y_N$ is a particular realisation) within stratum and between strata. We aim to explore situations where such assumptions may prove to be too restrictive. Although, it may still often be the case that the order in which units are chosen is not important. This leads to a natural generalisation to exchangeable sequences.

An infinite sequence $Y_1,Y_2,\ldots$ is said to be exchangeable if

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = f_{Y_1,\ldots,Y_n}(y_{\pi(1)},\ldots,y_{\pi(n)}) ,$$

for any subset of $Y_1,Y_2,\ldots$ and any $\pi\in\Pi$, the set of all finite permutations on $\{1,\ldots,n\}$ (Kingman, 1978). The assumption of exchangeability is characterised by a representation theorem, which states that there exists a conditional model $f_{Y|\theta}(y|\theta)$, where $\theta\in\Theta$ is the limit of some function of y_i's as $n\to\infty$, such that

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \int_\Theta \prod_{i=1}^{n} f_{Y|\theta}(y_i|\theta) f_\theta(\theta) d\theta ,$$

where $f_\theta(\theta)$ represents some prior belief for θ (Bernardo, 1996). This implies that an infinite exchangeable sequence is a mixture of IID sequences, or, in other words, it is conditionally IID, given the underlying distributional form. Diaconis (1977) further showed that this result is also approximately true for finite exchangeable sequences that are extendable (to a large exchangeable sequence, of size k say), with the error going to zero, at a rate like $1/k$.

The above representation can be generalised to multiple sequences, where each sequence is considered to be exchangeable (i.e., the sequences are partially exchangeable). Suppose that we can categorise a sequence $Y_1,Y_2,\ldots$ into H disjoint exchangeable subsequences and let $\mathbf{Y}_h$ denote a finite subset of those Y_i's that are in subsequence h (with the index subset denoted by S_h). Then, if $\mathbf{y}_h$ is a realisation of $\mathbf{Y}_h$, we have the following representation

$$f_{\mathbf{Y}_1,\ldots,\mathbf{Y}_H}(\mathbf{y}_1,\ldots,\mathbf{y}_H) = \int_\Theta \prod_{h=1}^{H}\prod_{i\in S_h} f_{Y|\theta_h}(y_i|\theta_h) f_{\theta_1,\ldots,\theta_H}(\theta_1,\ldots,\theta_H) d\theta_1\ldots d\theta_H , \qquad (1)$$

where θ_h is the set of underlying parameters associated with sequence h. If we further set $|S_h| = m_h$, then we have a scenario analogous to the two-phase stratified sampling in Section 2. Here, we consider the sequence of observations in individual strata to be exchangeable and dependencies across strata are characterised by the joint distribution $f_{\theta_1,\ldots,\theta_H}(\theta_1,\ldots,\theta_H)$ of the underlying parameter sets.

Under the above assumptions, it remains mathematically feasible to use the estimator $\bar{t}_{str}$ for $\bar{y}_U$. However, the calculation and estimation of $V(\bar{t}_{str})$ may become more cumbersome. Let Z_i be the indicator variable on unit i being selected for the first phase sample and $\mathbf{Z} = (Z_1, \ldots, Z_N)$. Consequently,

$$\begin{aligned} V(\bar{t}_{str}) &= V(E[\bar{t}_{str}|\mathbf{Z}]) + E(V[\bar{t}_{str}|\mathbf{Z}]) \\ &= V(\bar{t}^{(1)}) + E\left(V\left[\frac{1}{n'}\sum_{h=1}^{H} n_h \bar{y}_h|\mathbf{Z}\right]\right) \\ &= V(\bar{t}^{(1)}) + E\left(\sum_{h=1}^{H}\left(\frac{n_h}{n'}\right)^2 V[\bar{y}_h|\mathbf{Z}] + 2\sum_{a<b}\frac{n_a n_b}{(n')^2} Cov(\bar{y}_a, \bar{y}_b|\mathbf{Z})\right), \end{aligned} \tag{2}$$

where $\bar{t}^{(1)}$ is the sample mean from phase 1, assuming we know y_i for all $i \in S'$. The first term is the variance resulted from phase 1 sampling and the second term is the additional variance resulted from the subsampling in phase 2. Now, assuming $Y_1, \ldots, Y_N$ are still identically distributed with mean μ and variance σ^2, we can write the first term in expression (2) as

$$\begin{aligned} V(\bar{t}^{(1)}) &= E\left[(\bar{t}^{(1)} - \bar{y}_U)^2\right] \\ &= E\left[\left(\frac{1}{n'}\sum_{i\in S'} Y_i - \frac{1}{N}\sum_{i\in U} Y_i\right)^2\right] \\ &= E\left(\left(\frac{1}{n'} - \frac{1}{N}\right)\sum_{i\in S'} Y_i - \frac{1}{N}\sum_{i\notin S'} Y_i\right)^2 \\ &= E\left(\left(\frac{1}{n'} - \frac{1}{N}\right)\sum_{i\in S'} Y_i - \frac{1}{N}\sum_{i\notin S'} Y_i - \left(\frac{1}{n'} - \frac{1}{N}\right)n'\mu + \frac{1}{N}(N-n')\mu\right)^2 \\ &= E\left[\left(\frac{1}{n'} - \frac{1}{N}\right)^2\left(\sum_{i\in S'} Y_i - n'\mu\right)^2 + \left(\frac{1}{N}\right)^2\left(\sum_{i\notin S'} Y_i - (N-n')\mu\right)^2\right. \\ &\quad \left. -2\left(\frac{1}{n'} - \frac{1}{N}\right)\left(\frac{1}{N}\right)\left(\sum_{i\in S'} Y_i - n'\mu\right)\left(\sum_{i\notin S'} Y_i - (N-n')\mu\right)\right] \\ &= \left(\frac{1}{n'} - \frac{1}{N}\right)^2\left[n'\sigma^2 + 2\sum_{i,j\in S', i<j} Cov(Y_i, Y_j)\right] + \left(\frac{1}{N}\right)^2\left[(N-n')\sigma^2\right. \\ &\quad \left. +2\sum_{i,j\notin S', i<j} Cov(Y_i, Y_j)\right] - 2\left(\frac{1}{n'} - \frac{1}{N}\right)\left(\frac{1}{N}\right)\sum_{i\in S', j\notin S'} Cov(Y_i, Y_j)\,. \end{aligned}$$

The subsequent problem is in estimating the covariance terms

$$\sum_{i,j\in S', i<j} Cov(Y_i, Y_j)\,, \quad \sum_{i,j\notin S', i<j} Cov(Y_i, Y_j) \text{ and } \sum_{i\in S', j\notin S'} Cov(Y_i, Y_j)\,,$$

which incorporates covariances between Y_i's from the same stratum and across stratum. Now, for i and j in the same stratum, i.e., Y_i and Y_j are exchangeable, we may write

$$\rho_h := Cov(Y_i, Y_j) \approx V(E(Y_i|\theta)) = V(E(Y_i|F_{\mathbf{Y}_h}))\,,$$

where $F_{\mathbf{Y}_h}$ is the limiting empirical distribution of Y_i's in stratum h, if n_h is large and m_h/n_h is relatively small. We suggest estimating these within-stratum covariance terms using stationary bootstrapping (Politis and Romano, 1994) in each stratum. This is a generalisation to the standard bootstrapping, in which data are divided into blocks of random sizes (block sizes following a geometric distribution) and the blocks are re-sampled to form new samples. For simplicity, we also assume independence across strata (this can also be motivated practically when one agrees that changes in one stratum does effect others, or when such effects are considered minimal). Hence, $Cov(Y_i, Y_j) = 0$ for any pair i and j, from different strata. This will result in

$$\sum_{i,j\in S', i<j} Cov(Y_i, Y_j) \approx \sum_{h=1}^{H} \binom{n_h}{2} \rho_h$$

$$\sum_{i,j\notin S', i<j} Cov(Y_i, Y_j) \approx \sum_{h=1}^{H} \binom{\lceil n_h(N/n-1)\rceil}{2} \rho_h$$

$$\sum_{i\in S', j\notin S'} Cov(Y_i, Y_j) \approx \sum_{h=1}^{H} n_h(\lceil n_h(N/n-1)\rceil)\rho_h$$

where $\lceil n_h(N/n-1)\rceil$ is used to approximate $N_h - n_h$ and given that individuals in an exchangeable sequence behave similarly to each other (allowing us to approximate out-of-sample covariances with in-sample ones). We will also estimate σ^2 using the sample variance of all observed y.

The second term in (2), given independence across strata, is equal to

$$\tau := E\left(\sum_{h=1}^{H} \left(\frac{n_h}{n'}\right)^2 V[\bar{y}_h|\mathbf{Z}] \right).$$

This expectation is taken over all values of $\mathbf{Z}$ and cannot be evaluated given only one sample. Consequently, we propose estimating this expression again by using stationary bootstrapping. Although, the re-sampling here is taken over the union of S_h, i.e., m_h may change from re-sample to re-sample, and within each stratum of the re-sample (allowing the estimations of $V[\bar{y}_h|\mathbf{Z}]$). Within each re-sample, n_h is also estimated by $m_h n'/\sum m_h$.

4. An example

To implement our proposed methodology, we consider a practical example using the Australian AIDS survival data set[2]. In all steps where stationary bootstrapping is required, we set the bootstrapping parameter optimally to $p = (n^*)^{-1/3}$ (Politis and Romano, 1994), where n^* is the size of the sample we are re-sampling from, and the number of bootstrap samples is set to 1000.

The variable of interest y is the age (years) of patients at diagnosis. This is recorded for 2843 patients across Australia. An auxiliary variable x is readily available, which indicates the state of origin of each patient (NSW = New South Wales, QLD = Queensland, VIC = Victoria, Other = all other states). For our purpose here, let us assume this is our population and we aim to estimate $\bar{y}$, the

[2] Data by Australian National Centre in HIV Epidemiology and Clinical Research. Available in R package "MASS".

average age of those in the study of interest. However, we do not know the population stratum sizes N_h. Meanwhile, we undertake the judgment that $Y_1, Y_2, \ldots$ are independent across strata (states) and are exchangeable within stratum (which may not at all be an unreasonable judgment!).

We draw a sample S' using SRSWOR in phase 1 (and observe readings on x) and subsamples S_h are drawn from each strata in phase 2 using SRSWOR (and observe readings on y). A summary of the sample information is given in Table 1. The value of the corresponding two-phase stratified design estimator is given as $\bar{t}_{str} = 37.78714$, which can be compared to the true population mean $\bar{y} = 37.40907$. Sample variances seem to significantly vary across strata.

Table 1: Sample information for two-phase sampling on Australian AIDS survival data.

N	n'	h	n_h	m_h	$\bar{y}_h$	s_h^2	$\bar{t}_{str}$
2843	500	NSW	331	200	38.13	118.2142	37.78714
		QLD	40	27	37.7037	218.755	
		VIC	101	68	37.29412	90.30026	
		Other	28	19	35.63158	89.80117	

Table 2 records the estimated values for ρ_h and τ. The estimates for ρ_h are obtained through re-sampling within each stratum. The value for τ is obtained by both re-sampling the union of S_h and re-sampling within the resultant strata.

Table 2: Estimated variance and covariance using stationary bootstrapping.

h	ρ_h	τ
NSW	0.892956	0.3534294
QLD	3.464333	
VIC	1.074241	
Other	1.821655	

Table 3: Comparing sampling variance for $\bar{t}_{str}$.

Method/Assumption	Variance	Std. Dev.
Rao	0.0006403	0.02530365
Simulated	0.9817275	0.9908216
Exchangeable	0.5336707	0.7391561

The value of $\hat{V}(\bar{t}_{str})$ (and the corresponding standard deviation), under three different approaches, are presented in Table 3. The first estimate is obtained using the formula by Rao (1973), as given in Section 2. The second value is obtained from the population data, by re-calculating $\bar{t}_{str}$ repeatedly using random samples of size 500 and randomised phase 2 sampling ratio (all samples obtained using SRSWOR). This calculation is done for 10000 iterations and the sample variance of $\bar{t}_{str}$ across iterations is obtained. The formula by Rao (1973) clearly underestimates the variance of $\bar{t}_{str}$, due

to the assumption of IID observations. Meanwhile, our proposed approach, which caters for within stratum dependencies, produced an improved estimate for the variance (closer to the simulated variance from the population data).

5. Limitations to the method

There are several limitations to our approach that may be generalised or improved. Firstly, we have implemented a very Bayesian-unlike approach, in the sense that we did not specify a prior distribution for θ, nor a sampling distribution. More precisely, our method tries to capture the varying effect of θ through the bootstrapped samples. This is of course allowing the data to overtake any form of subjective prior information we may have for y, apart from the observed x values. Secondly, we have assumed independence across strata. Consequently, all covariances across strata were assumed to be zero. Relaxing this would again relate to specifying or estimating the joint behaviour between θ_h in expression (1). In addition, the example in Section 4 is based on a singular sample we have taken[3] and further simulation is required to observe the overall performance of our method. Further work should be done to compare our method to other general approaches to estimating variance in complex designs (Lohr, 2010).

6. Conclusion

In this paper, we considered a scenario of two-phase stratification sampling design, where observations within stratum are assumed to be exchangeable and strata are assumed to be mutually independent. A method is proposed for estimating the variance of the ratio estimator using stationary bootstrapping at various levels of the sampling procedure. An example considered here demonstrates that the standard variance estimate significantly overestimates the performance of the ratio estimator, while our method provided a more conservative approximation.

References

AHMED, M. S. (1998). A note on regression-type estimators using multiple auxiliary information. *Australian & New Zealand Journal of Statistics*, **40** (3), 373–376.

BERNARDO, J. M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, **Spec. II**, 111–121.

BHUSHAN, S., PANDEY, A., AND KATARA, S. (2008). A class of estimators in double sampling using two auxiliary variables. *Journal of Reliability and Statistical Studies*, **1** (1), 67–73.

COCHRAN, W. G. (1977). *Sampling Techniques*. 3rd edition. Wiley, New York.

DIACONIS, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese*, **36** (2), 271–281.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite population. *Journal of the Royal Statistical Society, Series B*, **31** (2), 195–233.

[3] Although we did test the method on a few other samples and have observed similar results.

HAMAD, N., HANIF, M., AND HAIDER, N. (2013). A regression type estimator with two auxiliary variables for two-phase sampling. *Open Journal of Statistics*, **3**, 74–78.

KINGMAN, J. F. C. (1978). Uses of exchangeability. *Annals of Probability*, **6** (2), 183–197.

LOHR, S. L. (2010). *Sampling: Design and Analysis*. 2nd edition. Brooks/Cole, Boston.

MUKERJEE, R., RAO, T. J., AND VIJAYAN, K. (1987). Regression type estimators using multiple auxiliary information. *Australian Journal of Statistics*, **29** (3), 244–254.

POLITIS, D. N. AND ROMANO, J. P. (1994). The stationary bootstrap. *Journal of American Statistical Association*, **89** (428), 1303–1313.

RAO, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **60** (1), 125–133.

SINGH, V. K., SINGH, G. N., AND SHUKLA, D. (1994). A class of chain ratio type estimators with two auxiliary variables under double sampling scheme. *Sankhyā, Series B*, **56** (2), 209–221.

TREDER, R. P. AND SEDRANSK, J. (1996). Bayesian sequential two-phase sampling. *Journal of the American Statistical Association*, **91** (434), 782–790.